# Ethical Review of M²SODAI Dataset

# 1. Potential Harms Caused by the Research Process

## 1.1. Research involving human subjects or participants:

**[Subject 1]** *Fair Wages*: all human research subjects or participants must receive appropriate compensation. If you make use of crowdsourcing or contract work for a particular task as part of your research project, you must respect the minimum hourly rate in the region where the work is carried out.

**[Answer 1]** This research was supported by a grant from Endowment Project of "Development of hyperspectral image analysis technology for rapid detection and identification of marine accidents based on machine learning approaches" funded by Korea Research Institute of Ships and Ocean engineering (PES4460). **All authors** received a three-year benefit from the fund for data collection, pre-processing, and experimental procedures.

**[Subject 2]** *Research involving human participants:* if the research presented involves direct interactions between the researchers and human participants or between a technical system and human participants, authors are required to follow existing protocols in their institutions (e.g. human subject research accreditation, IRB) and go through the relevant process. In cases when no formal process exists, they can undergo an equivalent informal process (e.g. via their peers or an internal ethics review).

**[Answer 2]** We hereby affirm and guarantee that the study presented does not encompass any direct interaction with human participants.

## 1.2. Data-related concerns:

The points listed below apply to all datasets used for submissions, both for publicly available data and internal datasets.

**[Subject 3]** *Privacy*: Datasets should minimize the exposure of any personally identifiable information, unless informed consent from those individuals is provided to do so.

**[Answer 3]** The proposed M²SODAI dataset is a RGB/HSI data taken at a high altitude of 1 km, and it was confirmed that personal sensitive information is not exposed in the presented dataset.

**[Subject 4]** *Consent*: Any paper that chooses to create a dataset with real data of real people should ask for the explicit consent of participants, or explain why they were unable to do so.

**[Answer 4]** This dataset is a dataset from 1 km in the sky, and images were taken over a wide range. There are few people in the dataset by photographing the ocean area and it is impossible to identify them. Since they are not selected participants, it is almost impossible to get individual signatures.

**[Subject 5]** *Deprecated datasets*: Authors should take care to confirm with dataset creators that a dataset is still available for use. Datasets taken down by the original author (ie. deemed obsolete, or otherwise discontinued), should no longer be used, unless it is for the purposes of audit or critical assessment. For some indication of known depreciated datasets, please refer to the NeurIPS list of deprecated datasets.

**[Answer 5]** We are the original creators of this dataset and guarantee that the dataset will be available in perpetuity.

**[Subject 6]** *Copyright and Fair Use:* While the norms of fair use and copyright in machine learning research are still evolving, authors must respect the terms of datasets that have defined licenses (e.g. CC 4.0, MIT, etc)

**[Answer 6]** The data set is provided by the authors, and this data is protected by the MIT lincense.

**[Subject 7]** *Representative evaluation practice*: When collecting new datasets or making decisions about which datasets to use, authors should assess and communicate the degree to which their datasets are representative of their intended population. Claims of diverse or universal representation should be substantiated by concrete evidence or examples.

**[Answer 7]** To construct a dataset that would be representative of the dataset's population (marine domain), we collected data from 12 different flight campaigns. However, due to the risk of flying in bad weather, data collected only in good weather is a limitation.

## 2. Societal Impact and Potential Harmful Consequences

Authors should transparently communicate the known or anticipated consequences of research: for instance via the paper checklist or a separate section in a submission.

The following specific areas are of particular concern:

**[Subject 8]** *Safety*: Contributors should consider whether there are foreseeable situations in which their technology can be used to harm, injure or kill people through its direct application, side effects, or potential misuse. We do not accept research whose primary goal is to increase the lethality of weapons systems.

**[Answer 8]** The main goal of our dataset is to pursue maritime safety through surveillance of maritime domains. As in the example in the Appendix, one of the big goals of our dataset is maritime rescue. For stability throughout the experiment, data in which a real person was in the water were excluded.

**[Subject 9]** *Security*: Researchers should consider whether there is a risk that applications could open security vulnerabilities or cause serious accidents when deployed in real world environments. If this is the case, they should take concrete steps to recommend or implement ways to protect against such security risks.

**[Answer 9]** This dataset collected data from areas other than sensitive areas such as military areas. Therefore, we are firmly convinced that the dataset will not pose a serious accident or security threat when distributed.

**[Subject 10]** *Discrimination*: Researchers should consider whether the technology they developed can be used to discriminate, exclude, or otherwise negatively impact people, including impacts on the provision of services such as healthcare, education or access to credit.

**[Answer 10]** These authors confirmed that the dataset they built did not have any negative social impact. For details, please refer to the discussion of the manuscript.

**[Subject 11]** *Surveillance*: Researchers should consult on local laws or legislation before collecting or analyzing any bulk surveillance data. Surveillance should not be used to predict protected categories, or be used in any way to endanger individual well-being.

**[Answer 11]** The data collection for this study was conducted in full compliance with local laws, and we hereby warrant that the safety of any individual has not been compromised.

**[Subject 12]** *Deception & Harassment*: Researchers should communicate about whether their approach could be used to facilitate deceptive interactions that would cause harm such as theft, fraud, or harassment, and whether it could be used to impersonate public figures and influence political processes, or as a tool to promote hate speech or abuse.

**[Answer 12]** In the process of collecting this data, we have confirmed that there is no connection to the topic in any way.

**[Subject 13]** *Environment*: Researchers should consider whether their research is going to negatively impact the environment by, e.g., promoting fossil fuel extraction, increasing societal consumption or producing substantial amounts of greenhouse gasses.

**[Answer 13]** The researchers considered that our dataset could facilitate future airplane research, stimulating fossil fuel extraction. However, this will be insignificant from a global perspective, and the need for a more eco-friendly drone data collection method was mentioned in the discussion part of the thesis.

**[Subject 14]** *Human Rights:* We prohibit circulation of any research work that builds upon or facilitates illegal activity, and we strongly discourage any work that could be used to deny people rights to privacy, speech, health, liberty, security, legal personhood, or freedom of conscience or religion.

**[Answer 14]** Our dataset is guaranteed to be free from problems under the laws of South Korea.

**[Subject 15]** *Bias and fairness:*  Contributors should consider any suspected biases or limitations to the scope of performance of models or the contents of datasets and inspect these to ascertain whether they encode, contain or exacerbate bias against people of a certain gender, race, sexuality, or other protected characteristics.

**[Answer 15]** Perhaps the only bias in the dataset is around the weather. However, this is not expected to have a negative social impact.

## 3. Impact Mitigation Measures

We propose some reflection and actions taken to mitigate potential harmful consequences from the research project.

**[Subject 16]** *Data and model documentation:* Researchers should communicate the details of the dataset or the model as part of their submissions via structured templates.

**[Answer 16]** This dataset is delivered as a structured template with examples through Github respository and Huggingface dataset.

**[Subject 17]** *Data and model licenses:* If releasing data or models, authors should also provide licenses for them. These should include the intended use and limitations of these artifacts, in order to prevent misuse or inappropriate use.

**[Answer 17]** This data is protected by the MIT lincense.

**[Subject 18]** *Secure and privacy-preserving data storage & distribution*: Authors should leverage privacy protocols, encryption and anonymization to reduce the risk of data leakage or theft. Stronger measures should be employed for more sensitive data (e.g., biometric or medical data).

**[Answer 18]** This dataset is not related to personal information protection/encryption/anonymization, and it is not biometric/medical data, so there is no risk of leakage.

**[Subject 19]** *Responsible release and publication strategy:* Models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, e.g. by requiring that users adhere to a code of conduct to access the model. Authors of papers exposing a security vulnerability in a system should follow the responsible disclosure procedures of the system owners.

**[Answer 19]** A possible misuse of the aircraft dataset would be the surveillance of filming data for national evidence areas. We have collected marine data on safe and open maritime domains.

**[Subject 20]** *Allowing access to research artifacts:* When releasing research artifacts, it is important to make accessible the information required to understand these artifacts (e.g. the code, execution environment versions, weights, and hyperparameters of systems) to enable external scrutiny and auditing.

**[Answer 20]** We plan to release all source code and datasets through Github.

**[Subject 21]** *Disclose essential elements for reproducibility:* Any work submitted to NeurIPS should be accompanied by the information sufficient for the reproduction of results described. This can include the code, data, model weights, and/or a description of the computational resources needed to train the proposed model or validate the results.

**[Answer 21]** We plan to release all source code and datasets through Github. In addition, all commends necessary for experiment reproduction are attached.

**[Subject 22]** *Ensure legal compliance:* Ensure adequate awareness of regional legal requirements. This can be done, for instance, by consulting with law school clinics specializing in intellectual property and technology issues. Additional information is required from authors where legal compliance could not be met due to human rights violations (e.g. freedom of expression, the right to work and education, bodily autonomy, etc.).

**[Answer 22]** We ensure adequate awareness of local legal requirements. We ensure adequate awareness of local legal requirements. For example, we guarantee that the intellectual property rights for the dataset belong to the authors and their affiliation (Korea Research Institude of Ships and Ocean Engineering).

Jonggyu Jang (signature)

Sangwoo Oh (signature)

Youjin Kim (signature)

Dongmin Seo (signature)

Youngchol Choi (signature)

Hyun Jong Yang (signature)