## **Name:** Gordon, Drew, Sameer

## **DataSheet For Coral\_Reef**

*“Highlighted all questions which cannot be answered”*

## **Introduction**

This data sheet template is taken directly from Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2018). Datasheets for datasets. *arXiv preprint arXiv:1803.09010*.

## **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.

Task was coral reef research and also coral reef segmentation for identifying these reefs.

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

● **Who funded the creation of the dataset?** If there is an associated grant, please provide the name of the grantor and the grant name and number.

● **Any other comments?**

## **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.

Aerial images of atolls, primarily from French Polynesia.

The dataset folder contains two subfolders named “JPG\_Labeled” and “JPG\_Unlabeled”.

The JPG\_Unlabeled folder contains raw images, as well as images after undergoing SSR, NIR, and SpecInd post processing.

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

1537 in the unlabeled folder, there are 4 variations of each atoll in the following formats: SSR, NIR, SpecInd

316 in the labeled folder. The atolls are labeled by a mask with different colors representing different features as follows: Yellow = land, Green = reef, Purple = Ocean

● **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 dataset is geographical…

● **What data does each instance consist of?** “Raw” data (e.g., unprocessed text or images) or features? In either case, please provide a description.

The data contains raw images, as well as images after post processing in the forms of SSR, NIR, and SpecInd. It additionally includes labeled versions of a subset of the images.

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

Labeling in order to differentiate between land, reef, and ocean will be associated with each instance.

● **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.

● **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.

The images each contain a different atoll. This means that the distinction between instances is made due to the atolls being described being unique to one another. Geographical relationships are not made clear.

● **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.

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

Some low quality images, some extremely high quality images could lead to discrepancies when a model receives an input at either extreme. Within the unlabeled dataset, each atoll has four versions of the same image with different processing.

● **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.

This 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.

## **Collection Process**

● **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.

● **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?

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

● **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)?**

● **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.

Past 2 years, can get more specific with interaction with dataset owner

● **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.

## **Pre-processing/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.

Data has been preprocessed in two ways. The unlabeled data has 4 formats, one with RGB, NIR, SpecInd, and SSR. These are image post processing formats which could change the interpretation of data based on which is used. The labeled data is processed due to the fact it is a mask of a real atoll. We are not given at least some of the original images from the labeled data. The labeling is based on the surface type depicted in images. Green = Reef, Yellow = Land, Purple = Ocean

● **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.

● **Is the software used to preprocess/clean/label the instances available?** If so, please provide a link or other access point.

● **Any other comments?**

## **Uses**

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

No.

● **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.

N/A

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

Land segmentation, ocean segmentation, environmental protection or monitoring.

● **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?

Low quality images could hinder use

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

● **Any other comments?**

## **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.

● **How will the dataset be distributed (e.g., tarball on website, API, GitHub)?** Does the dataset have a digital object identifier (DOI)?

● **When will the dataset be distributed?**

● **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.

● **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**

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

Email address: acortiz@colby.edu

● **Is there an erratum?** If so, please provide a link or other access point.

N/A

● **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)?

● **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 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.

● **Any other comments?**