Nacala-Roof-Material: Datasheet

Venkanna Babu Guthula[1], Stefan Oehmcke[1], Remigio Chilaule[2,3], Hui Zhang[1], Nico Lang[1], Ankit Kariryaa[1], Johan Mottelson[2], Christian Igel[1]

[1]University of Copenhagen
[2]Royal Danish Academy
[3]Mozambican NGO #MapeandoMeuBairro

A Motivation

A.1 For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled?

The dataset was created to support research on multi-task computer vision problems and to support mosquito-borne disease risk assessment in African cities. The list of tasks include classification, semantic segmentation, and instance segmentation of roofs and their material. While these tasks are closely related, each serves a different purpose and accurate segmentation of objects need not imply accurate object separation, and vice versa. The dataset is ideal for bench-marking methods for the above mentioned tasks.

A.2 Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The dataset was created by a team of researchers from University of Copenhagen and Royal Danish Academy. The drone imagery and building footprints were captured by Mozambican NGO #MapeandoMeuBairro. Map data was sourced from OpenStreetMap. The imagery and the building footprints were fused, re-registered, cleaned, verified, and split into given datasets by the authors from the University of Copenhagen and Royal Danish Academy.

A.3 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

The creation of the dataset is supported by the grant "Risk-Assessment of Vector-Borne Diseases Based on Deep Learning and Remote Sensing" by the Novo Nordisk Foundation. The grant number is NNF21OC0069116. The drone imagery used and building footprints were captured by #MapeandoMeuBairro under a development project supported by the Nacala Municipal Council, Mozambique and the ACRA Foundation, Italy.

A.4 Any other comment?

None.

B Composition

B.1 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

The dataset comprises of very-high resolution orthophotos captured through a drone and expert drawn polygons for all buildings with annotation of their roof material. The dataset covers three informal settlements in Nacala. Five classes of roof material are identified: metal sheet, thatch, asbestos, concrete, and no-roof. An example of a portion of the orthophoto and roof labels is shown in Fig. 1.

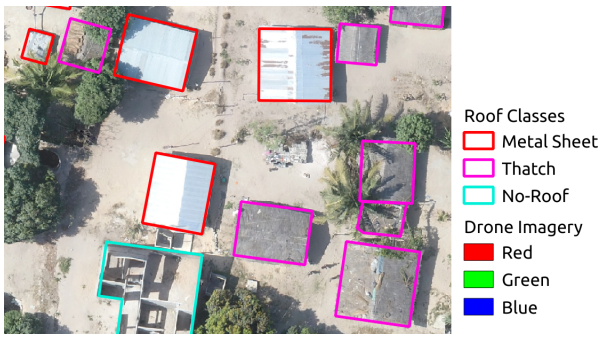

Figure 1: Drone Imagery with RGB (Red, Green, and Blue channels) and annotations

B.2 How many instances are there in total (of each type, if appropriate)?

The total number of building polygons in the data is 17954. The distribution of roof material classes is imbalanced. The number of buildings belonging to metal sheet, thatch, asbestos, concrete, and no-roof classes are 9776, 6428, 566, 174, and 1010, respectively.

B.3 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 contains all available instances. All informal settlements in Nacala that have drone orthophotos available are prepared as a dataset. Furthermore, all buildings visible in the orthophotos are included, and the five identified building classes cover all possible roof materials in the area, and the most predominant roof materials present in the wider Nacala region.

B.4 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 consists of aerial images and corresponding labels. Labels are building footprints with the attribute of roof class. The raw images are GeoTiff images tagged with a spatial reference system. The raw labels are GeoJSON files with the same spatial reference system as images.

B.5 Is there a label or target associated with each instance? If so, please provide a description.

The labels on the image are polygons describing the geometry of the building footprints and their associated roof material classes, as described above. In the raw data, the material class is saved under the attribute name of **mater_id** in GeoJSON files. The values of metal sheet, thatch, asbestos, concrete, and no-roof in the attribute are 1, 2, 3, 4, and 5, respectively. The same values are assigned to the patch labels.

B.6 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.

Everything is included. No data is missing.

B.7 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, the geometry and material attribute of each building footprint is independently recorded.

B.8 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.

The roof material classes are not balanced and are not geographically distributed uniformly. The data is split into training, validation, and test sets using stratified random sampling to account for the class imbalance. We created a square grid of 225 meters and counted the roof types in these cells. Then we partitioned the cells into three sets based on the class counts to achieve a similar class distribution in each dataset, where we prioritized the distribution of minority classes (i.e., concrete and asbestos). We defined that a building only belongs to a specific grid cell if its centroid falls into the cell. These grid cells separate the images and labels into training, validation and test sets. See Fig. 2 as an example. Initially, only two informal settlements were labelled and therefore, only these two settlements are divided into the 3 sets. The third informal settlement was labelled later and treated as a second test set.

B.9 Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

The images exhibit a high level of details and the building footprint geometry and material attributes are meticulous noted by experts. The dataset is free from errors, noise, or redundancies to the greatest extent possible but we acknowledge that even with expert craftsmanship, there is always a chance of human error.

B.10 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 entirely self-contained.

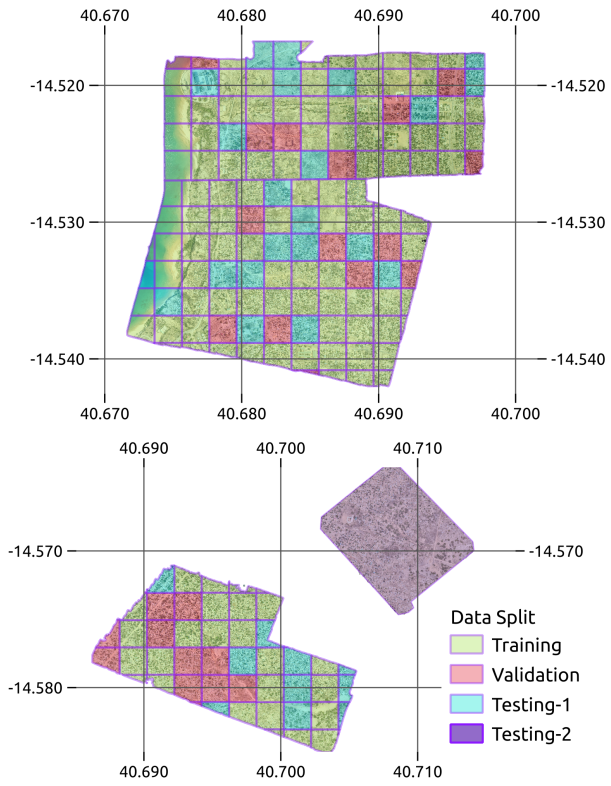

Figure 2: Visualisation of the training, validation and testing sets with reference to longitude and latitude

B.11 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' nonpublic communications)? If so, please provide a description.

The dataset does not contain any confidential data.

B.12 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.

B.13 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.

B.14 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.

It is not possible to identify individuals in the drone imagery. In any publicly available and geo-coded image, it is possible to identify individual houses and reverse geocode into a human-readable address.

B.15 Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race 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.

B.16 Any other comments?

None.

C Collection Process

C.1 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 data is observable as images. The ground sampling distance of pixel or spatial resolution of the imagery is ≈ 4.4 cm/pixel. QGIS was used for the visualization of images and re-registration, cleaning, and verification of building footprints and their attributes [1].

C.2 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 drone imagery was captured using a DJI Phantom 4 Pro drone and processed using AgiSoft Metashape software [2]. The building footprints were updated on OpenStreetMap [3] and downloaded from OpenStreetMap under the Open Database License (ODbL). The missing labels and all geometric and attribute errors were corrected in QGIS software.

C.3 If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

The data was prepared based on its availability. The original project, supported by the Nacala Municipal Council, Mozambique, studied the risk of erosion in informal settlements. The areas selected to be mapped with drones are those where erosion poses the highest risk to physical property and human life. All data from that project was made available and used in this dataset.

C.4 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 drone imagery was captured by Mozambican NGO #MapeandoMeuBairro. Nacala residents and local university students performed the field data collection, receiving stipends and data bundles. The polygons and attributes of building footprints were corrected by authors from the University of Copenhagen and the Royal Danish Academy as described above.

C.5 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.

All drone imagery was captured between October and December 2021. All labels are manually annotated on the imagery beginning January 2022 until May 2024.

C.6 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.

No ethical review was conducted. All data collection, including drone flights and on-site mapping, was approved and led by the Nacala Municipal Council and facilitated on the ground by neighbourhood-level authorities.

C.7 Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

The data was not collected from individuals.

C.8 Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

N/A.

C.9 Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

N/A.

C.10 If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

N/A.

C.11 Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A.

C.12 Any other comments?

None.

D Preprocessing/Cleaning/Labeling

D.1 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.

Because of the large size of raw aerial imagery, the images of training and validation sets were cropped to 512×512 pixels of **patches**. The data processing optimized is usefulness in training deep learning models. The total number of patches in the training and validation sets are 8366 and 1799, respectively, after cropping them without any overlap. The test sets were provided without cropping into patches, so these images are provided in different sizes. The images

in test sets are not cropped because dividing them into patches may lead to under- or over-estimation of instances and may influence counting accuracy over large areas. The test-1 set consists of 22 images and the test-2 set consists of a single image. There are 10930, 2956, 2527 and 1541 buildings in train, validation, test-1 and test-2 sets, respectively.

For the classification of buildings into different roof classes, the DINOv2 features were extracted for train, validation and test-1 sets and made available with the dataset. These features of the train, validation, and test-1 buildings are saved in train.npy, test.npy and valid.npy files, respectively. Each row in the NumPy file is a DINOv2 feature of a single building along with a label in its last column. The five roof classes are there in the data: 1-Metal Sheet, 2-Thatch, 3-Asbestos, 4-Concrete and 5-No Roof. The feature extraction is further explained in the in our research paper that was submitted to the NeurIPS 2024 Datasets and Benchmarks Track.

D.2 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.

Yes. Along with patches and their labels, the dataset contains the raw data.

D.3 Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

The Python script is used to prepare patches and label different deep-learning models (e.g., UNet and YOLOv8). The Python script is available in our GitHub repository. Repository link: https://github.com/mosquito-risk/Nacala

D.4 Any other comments?

None.

E Uses

E.1 Has the dataset been used for any tasks already? If so, please provide a description

At the time of preparing this datasheet, the dataset was only used for tasks performed in the paper that was submitted to NeurIPS Datasets and Benchmarks Track 2024.

E.2 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.

E.3 What (other) tasks could the dataset be used for?

There are other objects in the images that can also be mapped, for example, trees, roads, water bodies, etc.

E.4 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 dataset consumer 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 risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?

There is no risk of using this dataset.

E.5 Are there tasks for which the dataset should not be used? If so, please provide a description.

None.

E.6 Any other comments

None.

F Distribution

F.1 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 is publicly available on the internet.

F.2 How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

The data and code are available on ERDA (https://sid.erda.dk/sharelink/aHw1Pey5BC) and GitHub (https://github.com/mosquito-risk/Nacala), respectively.

F.3 When will the dataset be distributed?

The dataset was first released in June 2024.

F.4 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 dataset is released under Open Data Commons Open Database License (ODbL) v1.0 licence.

F.5 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

F.6 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

F.7 Any other comments?

None.

G Dataset Maintenance

G.1 Who is supporting/hosting/maintaining the dataset?

Venkanna Babu Guthula maintains the preprocessed data with splits in ERDA, and Remígio Chilaule maintains the raw data.

G.2 How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

Curators of this dataset can be communicated through our GitHub repository issues.

G.3 Is there an erratum? If so, please provide a link or other access point.

No. This datasheet was prepared with our first release.

G.4 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 dataset consumers (e.g., mailing list, GitHub)?

In case of any updates, we will communicate through our GitHub repository.

G.5 If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the 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.

N/A.

G.6 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 dataset consumers. The dataset has already been updated; older versions are kept around for consistency

N/A.

G.7 If others want to extend/augment/build on this dataset, is there a mechanism for them to do so? If so, is there a process for tracking/assessing the quality of those contributions. What is the process for communicating/distributing these contributions to users?

N/A

G.8 Any other comments?

None.

References

[1] QGIS. `https://qgis.org/en/site/`, 2024. (Accessed on 05/22/2024).

[2] AgiSoft Metashape. `https://www.agisoft.com/`, 2024. (Accessed on 05/22/2024).

[3] OpenStreetMap. `https://www.openstreetmap.org/`, 2024. (Accessed on 05/22/2024).