MOTIVATION.

[question1] 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.
[answer1] The dataset was created for Classifying neighbourhoods in South Africa According to 4 neighbourhood
types: wealthy (a combination of suburbs, smallholdings, farms), non wealthy (combination of townships, informal areas, collective living quarters, villages), non residential areas (combination of i industrial areas, commercial lands, parks and recreational areas, vacant land) and background (all land that does not have a building on it). The disaggregated labels (12 classes rather than 4) are also available with the dataset. The specific application the authors created this dataset for is to enable researchers and policymakers to quantify the effects of spatial apartheid over time, for the specific purpose of helping to uncover and working to reverse its effects.

[question2] Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?
[answer2] Raesetje Sefala (University of the Witwatersrand), Timnit Gebru (DAIR), Nyalleng Moorosi (Google) and
Richard Klein (University of the Witwatersrand) created this dataset. The project was conceived in 2017 when Nyalleng Moorosi was at the Council for Scientific and Industrial Research (CSIR), South Africa and Timnit Gebru was at Microsoft Research in the USA. The dataset was created using a combination of other pre-existing datasets: The Enumeration Areas (EAs) dataset created in 2011 by Statistics South Africa (Stats SA)--a government agency responsible for conducting the census; Geographically referenced
(Geo-referenced) buildings dataset created by Eskom (a South African electricity public utility company) in partnership with the Council for Scientific and Industrial Research (CSIR) consisting of building count data in South Africa from 2006 to 2016; and satellite images from 2006-2017 from the South African National Space Agency (SANSA).

[question3] 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.
[answer3] Funding was provided through five sources: The South African Department of Science and Technology and
CSIR (as a masters scholarship award to Raesetje Sefala), Google (research award to Raesetje Sefala and compute credits), the Deep learning Indaba and Nvidia (Nvidia Titan V prize for best poster presentation at the 2018 Deep Learning Indaba summer school), and the DAIR institute.

[question4] Any other comments?
[answer4] No

COMPOSITION.

[question5] What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.
[answer5] Each instance is a 256x256 satellite image of South Africa from 2011 paired with masks of the

neighbourhood type each building cluster represents. We also include satellite images from other years (2006-2017) without masks. The 256x256 images were obtained by tiling each high-resolution satellite image for ease of processing. Images are taken from the SPOT sensor with varying resolutions in different years: one pixel represents 10 meters on the ground for 2011 and each satellite image consists of 21,688 x 21,688 pixels. Before tiling the satellite images by images of 256x256, we upsampled them to 21,760x21,760 using the GDAL library (https://gdal.org/). This was done in order to make sure that a single satellite image can be fully covered by an integer number of 256x256 images.
Table 1 shows the resolution of satellite images for each year and the number of images per year. There are two sets of masks, one set has 12 neighbourhood classes and the other set has 4 neighbourhood classes. Five example instances are shown in figure 1 below. The 12 and 4 classes are listed in answering question 1.

[question6] How many instances are there in total (of each type, if appropriate)?

[answer6] There are 3,973,750 256x256 satellite images in total from 20211 with associated masks. This
corresponds to the 550 satellite images in total, which were originally of resolution 21,688X21,688 and which we upsampled to 21,760x21,760. We also include satellite images from 2006-2017 without labels (a total of 6,218 satellite images).

[question7] 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 (for example, 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 (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).

[answer7] The dataset is not sampled (represents all images of South Africa in 2011). However, given the highly
imbalanced nature, and a large amount of vacant land, we perform experiments on a sampled version of our dataset (see associated paper for details).

[question8] What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.

[answer8] Each instance consists of a 256x256 subset of a raw satellite image, a 12-class mask of features and a
4-class mask of features. The satellite images have 3 channels (RGB) and were taken using the SPOT5 Satellite sensor. The 12-class masks are PNG images with 3 channels (RGB), with 12 distinct colours to represent the 12 individual classes (e.g. green[0,255,0] - farm, yellow[255,255,0] - township, white[255,255,255] - background). The 4 class masks are also PNG images with 3 channels and 4 distinct colours to represent the 4 individual classes.

[question9] Is there a label or target associated with each instance? If so, please provide a description.

[answer9] For each satellite image, corresponding masks are labelled at the pixel level to represent the
neighbourhood type for each pixel on the satellite image.

[question10] Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example,

because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.

[answer10] Everything is included. However, if there are missing instances in our source datasets due to errors (e.g.
building dataset), our dataset will also miss these instances.

[question11] Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

[answer11] Each instance is associated with a latitude and longitude. This allows us to know which physical location
each instance represents, and how instances are spatially related to each other.

[question12] Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

[answer12] We have performed experiments using a subset of our data and report the train/validation/test split we
used in these experiments as part of the dataset metadata. What splits people should use depends on the specific task/application they work on. We recommend ensuring that all images representing the same neighbourhood are in the same split.

[question13] Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

[answer13] Yes.

- The masks are created in such a way that each house is represented using a circle with a diameter of 0.0007 decimal degrees irrespective of the type/size of the building. We did that because we did not have a source that captures the exact sizes of the buildings throughout the country consistently across all neighbourhood types (available datasets do not capture informal settlements and some villages consistently). The pink polygon in figure 2 represents an industrial building on the satellite image, the representation is not a tight bound around the extent of the building but instead a circular polygon of diameter 0.0007 decimal degrees over the centroid of the building. Similarly, that is how we created the rest of the dataset.
- The building dataset has other potential sources of error especially around the labelling of informal settlements. Individual buildings in these neighbourhoods are usually very difficult to distinguish from satellite images. Small buildings camouflaged by the surrounding environment can also be difficult to detect.
- Although we have taken steps to verify that the information is correct, another potential source of error is how we labelled the townships in our dataset. We followed the procedure in section 3 of our associated paper to distinguish suburbs from townships, given that they are both labelled as formal residential neighbourhoods in the EA dataset. As noted in the paper, our labeling process could result in some townships that are labeled as suburbs or vice versa. It is much more likely that we have misclassified some townships as suburbs as Wikipedia may not have all the labels of townships in 2011 and if 2 people label something as a township and we cannot find the name listed as a township in other sources, we classify it as a suburb.
- Additionally, the manner in which we demarcated wealthy and non wealthy neighborhoods can be a source of error. Some collective living quarters that are close to places of economic activity may be wealthy neighborhoods. In addition, while townships were allocated very low budgets during apartheid, there are now wealthy households within townships.

[question14] Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, 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 (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

[answer14] Yes, the dataset is self-contained. Although users can substitute the satellite images for other types
(more/fewer channels)/pixel resolutions. They will have to match the geographical referencing so that the masks can align properly. The satellite images included in our dataset are from the South African National Space Agency, and they have permitted us to release the dataset for research purposes only. We have similar permission from Eskom to release the building count dataset for research purposes. The EA dataset is publicly available as part of the South African census data.

[question15] Does the dataset contain data that might be considered confidential (for example, 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.

[answer15] Our dataset does not include confidential data and only provides masks at the neighbourhood level.
However, the building count dataset which is used in the construction of our data locates each building in South Africa. If an individual lives in a particular building and does not want their house/building to be located with this building dataset, it is unclear if they can do so.

[question16] Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

[answer16] No, it does not.

If the dataset does not relate to people, you may skip the remaining questions in this section.

[SkipA] NO

[question17] Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

[answer17] The dataset indirectly identifies subpopulations by neighbourhood types which can roughly approximate
standards of living. The labels were created by Statistics South Africa (a government entity responsible for the census) for the census. We overlaid their dataset with ours so that we can label the buildings according to these neighbourhood types.

[question18] Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.

[answer18] No, it is not possible to identify individuals. This dataset labels clusters of buildings according to their
type.

[question19] Does the dataset contain data that might be considered sensitive in any way (for example, 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.

[answer19] The enumeration area (EA) dataset from the census, which is public, contains demographic data for each
Main Place (A Main place is a group of EAs) but we do not include this data as part of the data we are releasing, it was released as part of the 2011 South African census. While the EA dataset already labels EAs by associating them with 12 types of neighborhoods according to the land's intended use, our dataset further approximates the location of building clusters within the EA polygons to distinguish between intended and actual land use.

[question20] Any other comments?

[answer20] No

COLLECTION PROCESS.

[question21] How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

[answer21] The dataset was created using 3 other datasets as input.

- Satellite images: Captured using the SPOT5 Satellite sensor.
- Enumeration Area dataset: Directly observed and captured during the time they made the 2011 South African census
- Building dataset: systematically collected and cleaned by a group of subjects, this dataset went through various stages of verification, more information about the dataset can be found at https://www.ee.co.za/wp-content/uploads/legacy/PositionIT%202009/PositionIT%202010/SPOT.pdf

Building clusters were inferred using a buffing algorithm of diameter 0.007 degrees. This process can introduce noise as mentioned in answering question 13.

[question22] What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?

[answer22] Satellite images: Captured using the SPOT5 Satellite sensor.

- Enumeration Area dataset: Unknown

- Building dataset: Unknown

- In our dataset processing phase, we used the QGIS software for tasks requiring spatial processing (projections, overlays etc).

[question23] If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?

[answer23] The sample is of satellite images of all of South Africa.

[question24] Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?

[answer24] Source datasets:

- Satellite images: The South African National Space Agency bought the images from a third-party source

- Enumeration Area dataset: We are not sure

- Building dataset: Contractors, we do not know if they were compensated.

In addition to those listed in question 2, we recruited 10 volunteer students who grew up in townships to
aid in labeling townships.

[question25] Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

[answer25] Source datasets were collected during these timeframes:

- Satellite images: 2006--2017

- Enumeration Area dataset: The labels are for 2011 but the census verification process lasted until 2013.

- Building dataset: 2006-2017

Our ground truth data creation process was done between 2018 and 2020.

[question26] Were any ethical review processes conducted (for example, 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.

[answer26] There were no review processes conducted by a review board. However, see answer to Q29 on an
analysis of potential risks and harms.

If the dataset does not relate to people, you may skip the remaining questions in this section.

[SkipB] NO

[question27] Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?

[answer27] The dataset was obtained via other sources listed in question 2 (South African Space Agency, Eskom, and
Statistics South Africa).

[question28] 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.

[answer28] For the following source datasets used in our annotations,

- Satellite images: No: the data was collected using satellites.
- Enumeration Area dataset: Created under the South African 2011 census project.
- Building dataset: To our knowledge building occupants were not notified that a building dataset of all buildings in South Africa was constructed and that the building they occupy is in the dataset.

[question29] 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.

[answer29] Yes, for the EA dataset as this was done under the South African census project and is publicly available
data. Stats SA is mandated to provide the state with information about the economic, demographic, social and environmental situation in the country. This is in line with the Statistics Act, (Act No. 6 of 1999), and the fundamental principles of official statistics of the United Nations. Legally, Section 16 of the Statistics Act (Act 6 of 1999) obliges a respondent to answer all questions put to them by an officer of Statistics South Africa. Section 17 of the Statistics Act guarantees the confidentiality of your information. The data collected is used for statistical purposes only and no-one can access data on an individual level.

[question30] 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).

[answer30] See answer for Question 27. Beyond that we do not know if individuals have a mechanism to revoke
their consent for the collection of the EA dataset. The data collected is used for statistical purposes only and no-one can access data on an individual level.

[question31] Has an analysis of the potential impact of the dataset and its use on data subjects (for example, 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.

[answer31] Our analysis consisted of speaking to various stakeholders and incorporating their feedback. Some of the
researchers working on this dataset also grew up in townships and have seen various manners in which data driven systems can marginalize people in South Africa. This led us to believe that our dataset should
only be available for research, rather than commercial, purposes. Our dataset is also only available

through requests, rather than on a website where it can be downloaded by anyone. This allows us to grant access only to those who request it for uses endorsed by us.

[question32] Any other comments?

[answer32] No

PREPROCESSING/CLEANING/LABELING.

[question33] Was any preprocessing/cleaning/labeling of the data done (for example, 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 remaining questions in this section.

[answer33] Our associated paper and supplementary materials describe in detail how the dataset was acquired and
processed. In short, we started with centroids of building locations, polygons with labels denoting land use as mandated by the government, and satellite images of South Africa, we performed the following steps;

- We inflated the building centroids into polygons as shown in figure 3 and 4 to cover the houses.
- Intersect the inflated building polygon data with the polygons denoting land use
- And smoothed overlapping building polygons by neighbourhood type

Given the way we created the dataset, any building that was not captured (missing data) in the building dataset will also
not be represented in our dataset.

[question34] Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

[answer34] Yes. Raw data from all sources (building count data, satellite images, EA dataset, has been saved).

[question35] Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.

[answer35] We are making the code we used to process the data available with the dataset.

[question36] Any other comments?

[answer36] No

USES.

[question37] Has the dataset been used for any tasks already? If so, please provide a description.

[answer37] The dataset has been used to perform experiments in section 5 of the associated paper related to
neighborhood classification in South Africa including:

- Training a model on 8 provinces and testing on the 9th to investigate the visual similarity between

provinces.

- Investigating if our dataset can be used to detect the evolution of neighborhoods in South Africa between 2011 and 2017. For instance, what types of neighborhoods were built on vacant land by 2017 that did not exist in 2011?

[question38] 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.

[answer38] We plan to create such a repository and reach out to people who request the dataset to update us with the paper/task they have used it for. We will update the datasheet with the location of the repository.

[question39] What (other) tasks could the dataset be used for?

[answer39] The dataset can be used to experiment with semantic segmentation more generally. It can also be merged with other existing datasets to make inferences about the standard of living in various South African neighborhoods, and the characteristics of people who live in these neighborhoods (using census data). Insurance, bank and other types of companies have, in the past, used these types of datasets to help them make predictions about the type of loans people can receive or the types of insurance groups can receive. Many of these practices have been found to be discriminatory so we do not allow for our dataset to be used in those scenarios.

[question40] 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 (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?

[answer40] There are entities that discriminate against people based on their zip codes in many countries including South Africa. Also see answers to question 39 below.

[question41] Are there tasks for which the dataset should not be used? If so, please provide a description.

[answer41] While our dataset does not identify individuals, we know that there are cases of entities using information about neighborhoods, and linking them to data they have on individuals to make inferences about them in ways that are discriminatory. We plan to screen for uses of our dataset by asking people to fill out a request form to obtain it, including what they plan to do with the dataset, asking for an update on what the dataset was used for and tracking it in our repository. We will not accept use cases that:

- Enable harassment, threatening, intimidating, predatory or stalking conduct;
- Determine financial consequences such as interest rates, insurance prices or loans;
- Pertain to the military or aid in drone targeting;
- Have commercial deployment. This dataset is to be used for strictly research purposes.

[question42] Any other comments?

[answer42] No

DISTRIBUTION.

[question43] Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.
[answer43] The dataset will only be available for academic research use. It will be available via this request form.
[question44] How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?
[answer44] The dataset will be hosted on a Google Cloud Platform in a Bucket and will be available based on
requests on https://forms.gle/x6YmS96VVPgsUSiQ6. The dataset will be associated with a DOI upon release which will be added to the datasheet.
[question45] When will the dataset be distributed?
[answer45] The dataset will be available for release on December 1, 2021.
[question46] 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.
[answer46] This dataset is freely available for academic and non-academic entities to use for non-commercial
purposes such as academic research, teaching, scientific publications, or personal experimentation. Permission is granted to use the data given that users agree the terms below.
1. Users should include a reference to the dataset in any work that makes use of the dataset. For research papers, cite our preferred publication as listed on our website; for other media cite our preferred publication as listed on our website or link to the dataset website.
2. Subject to compliance with these terms, users are granted a limited, non-exclusive, non-transferable, non-sublicensable, revocable license to access and use the dataset.
3. Users should not distribute this dataset or modified versions. It is permissible to distribute derivative works in as far as they are abstract representations of this dataset (such as models trained on it or additional annotations that do not directly include any of our data), given that the use cases are in line with those listed in this datasheet.
4. The dataset or any derivative work may not be used for commercial or military purposes as, for example, licensing or selling the data, or using the data with a purpose to procure a commercial or military gain.
5. That all rights not expressly granted to the users are reserved by us (Authors).
[question47] 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.

[answer47] The South African National Space Agency and Eskom have given us permission to distribute the satellite
and building count datasets respectively, for research use.

[question48] 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.

[answer48] No.

[question49] Any other comments?

[answer49] No

MAINTENANCE.

[question50] Who will be supporting/hosting/maintaining the dataset?

[answer50] Raesetje Sefala is supporting/maintaining the dataset.

[question51] How can the owner/curator/manager of the dataset be contacted (for example, email address)?

[answer51] Email them at sa.spatialproject@gmail.com.

[question52] Is there an erratum? If so, please provide a link or other access point.

[answer52] This is the first version of the dataset release. Any changes/updates will be posted on the dataset's
official website (link to be added to the datasheet with official release). The changes/updates will also be emailed to those who received the dataset through an official datast request.

[question53] Will the dataset be updated (for example, 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 (for example, mailing list, GitHub)?

[answer53] If we do find errors or other information that should be corrected, we will update the dataset
accordingly and post the update on the dataset webpage as well as email registered users of the dataset.

[question54] If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, 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.

[answer54] N/A

[question55] 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.

[answer55] Older versions will be kept and maintained for consistency even if newer versions are released.

[question56] 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 dataset consumers? If so, please provide a description.

[answer56] Those who would like to make contributions can request to use the dataset like others and explain what
they plan to do. We plan to regularly poll users to update our repository of dataset use. If they release derivative datasets which require accessing our dataset, we require that our dataset still be accessed via our form so that we can track what it is used for.

[question57] Any other comments?

[answer57] No