Supplementary Material and Datasheet for the WorldStrat Dataset

J. Cornebise, I. Oršolić, F. Kalaitzis

2022-06-16

Contents

1 Downloading the Dataset and the Software Package 4

2 Cloud coverage statistics 4

3 Full List of Hyperparameters for Benchmark 4

4 Datasheet 6

Motivation 6
For what purpose was the dataset created? . . . 6
Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)? . . . 6
Who funded the creation of the dataset? . . . 6
Any other comments? . . . 7

Composition 7
What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? . . . 7
How many instances are there in total (of each type, if appropriate)? . . . 7
Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? . . . 8
What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? . . . 8
Is there a label or target associated with each instance? . . . 10
Is any information missing from individual instances? . . . 10
Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? . . . 10
Are there recommended data splits (e.g., training, development/validation, testing)? . 10
Are there any errors, sources of noise, or redundancies in the dataset? . . . 13
Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? . . . 13
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)? . . . 15
Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? . . . 15
Does the dataset relate to people? . . . 15

Does the dataset identify any subpopulations (e.g., by age, gender)? . . . 15
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? . . . 15
Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial 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)? . . . 16
Any other comments? . . . 16

Collection Process **16**
How was the data associated with each instance acquired? . . . 16
What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? . . . 16
If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)? . . . 16
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)? . . . 16
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)? . . . 17
Were any ethical review processes conducted (e.g., by an institutional review board)? 17
Does the dataset relate to people? . . . 18
Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)? . . . 18
Were the individuals in question notified about the data collection? . . . 18
Did the individuals in question consent to the collection and use of their data? . . . 18
If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? . . . 18
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? . . . 18
Any other comments? . . . 18

Preprocessing/cleaning/labeling **18**
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)? . . . 19
Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? . . . 19
Is the software used to preprocess/clean/label the instances available? . . . 19
Any other comments? . . . 19

Uses **19**
Has the dataset been used for any tasks already? . . . 19
Is there a repository that links to any or all papers or systems that use the dataset? . 19
What (other) tasks could the dataset be used for? . . . 19
Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? . . . 20
Are there tasks for which the dataset should not be used? . . . 20
Any other comments? . . . 20

Distribution **20**
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? 20
How will the dataset will be distributed (e.g., tarball on website, API, GitHub) 20
When will the dataset be distributed? 21
Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? 21
Have any third parties imposed IP-based or other restrictions on the data associated with the instances? 21
Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? 21
Any other comments? 21

Maintenance **21**
Who will be supporting/hosting/maintaining the dataset? 21
How can the owner/curator/manager of the dataset be contacted (e.g., email address)? 21
Is there an erratum? 22
Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? 22
If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? 22
Will older versions of the dataset continue to be supported/hosted/maintained? 22
If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? 22
Any other comments? 22

References **23**

List of Figures

1 Summarizing the construction and classes of the WorldStrat dataset. 9
2 Frequency of class occurrences within the dataset, for SMOD urban density (top left), IPCC land use (top right), LCCS land use (bottom). 12

List of Tables

1 Hyper parameters used in the Benchmark of Section 4 of Cornebise et al. [2022a] and their values. 5
2 Land use classes according to the LCSS and the IPCC classifications. LCCS comprises of 23 classes and 14 sub-classes. IPCC groups those into 6 coarser classes. 11
3 The GHSL-SMOD dataset comprises of 3 classes at level 1, and 8 sub-classes at level 2, as described in the GHSL Data Package. 13
4 Frequency of LCCS class occurrences in the dataset. 14
5 Frequency of SMOD (left) and IPCC (right) class occurrences in the dataset. . . 15

1 Downloading the Dataset and the Software Package

The dataset, along with its machine-readable metadata, is hosted on CERN-backed Zenodo data repository: https://zenodo.org/record/6810792 [Cornebise et al., 2022b]. Its long-term maintenance is discussed in the Datasheet.

The software package is available on GitHub at https://github.com/worldstrat/worldstrat. This includes reproducible code for the Benchmarks of Section 4 of [Cornebise et al., 2022a], following the ML Reproducibility Checklist [Pineau et al., 2021a,b].

The project also has its own website available at https://worldstrat.github.io/, containing links to the dataset and software package download and information on how to cite.

The authors hereby state that they bear all responsibility in case of violation of rights, etc., and confirm that the data license is as follows:

- The low-resolution imagery, labels, metadata, and pretrained models are released under Creative Commons with Attribution 4.0 International (CC BY 4.0)[1];
- The high-resolution imagery from Airbus is distributed under Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0)[2];
- The code is distributed under BSD license.

2 Cloud coverage statistics

The mean of the cloud coverage *over the Sentinel 2 product areas* is 7.98 %, with a standard deviation of 14.22. The quantiles are:

- 0.025: 0.00%
- 0.25: 0.00%
- 0.5: 0.66%
- 0.75: 10.05%
- 0.975: 49.95%

It is important to note that this cloud cover percentage, as mentioned in the article and datasheet, is calculated on the entire product size of the provider, which varies in size but is much larger than the 2.5km^2 we target.

This means that even an image with a large cloud cover percentage can be cloud free, and in extreme cases (though unlikely), vice-versa.

Also there are indeed considerable difference across sampled regions and land cover types. A simple example would be rainforests and non-desert equatorial regions. Using a strict no-cloud policy would make sampling enough low-resolution images either impossible or would make the temporal difference extremely large (up to 7 years for some AOIs).

With that in mind, we strived to keep the cloud coverage as low as possible, ideally under 5%, while maintaining the temporal difference as small as possible.

3 Full List of Hyperparameters for Benchmark

Table 1 lists all the hyperparameters employed in the Benchmark covered in Section 4 of [Cornebise et al., 2022a].

[1] https://creativecommons.org/licenses/by/4.0/
[2] https://creativecommons.org/licenses/by-nc/4.0/

Hyperparameter	Value
Batch size	48
Floating-point precision	16-bit
Learning rate	0.0001
Learning rate decay	0.97
Learning rate patience	3
Weight decay	0.0001
Hidden channels	128
Residual layers	1
Input size	160x160
Output size	500x500
Chip size	50x50
Kernel size	3
Super-resolution kernel size	1
Shift pixels by	2
Shift mode	Lanczos (sub-pixel)
Shift step	0.5
Weight MSE	0.3
Weight MAE	0.4
Weight SSIM	0.3
Random seeds	122938034, 431608443, 315114726
Dataset random seed	386564310
Number of revisits	8 for multi-frame, 1 for single-frame
Max epochs	15
Max steps	50000
Cosine Annealing T_0	300

Table 1: Hyper parameters used in the Benchmark of Section 4 of Cornebise et al. [2022a] and their values.

4 Datasheet

This Datasheet for Dataset follows the template from Gebru et al. [2021].

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.

Analyzing the planet at scale with satellite imagery and machine learning is a dream that has been constantly hindered by the cost of difficult-to-access highly-representative high-resolution imagery. We introduced this dataset to remediate this.

The aim was to create, with a reasonable budget, the largest and most varied such publicly available dataset. We wanted to provide as broad and application-agnostic a representation of the physical features of the world as possible, by curating nearly 10,000 km^2 of unique locations to ensure stratified representation of all types of land-use across the world: from agriculture to ice caps, from forests to multiple urbanization densities. We also enrich those with locations typically under-represented in ML datasets: sites of humanitarian interest, illegal mining sites, and settlements of persons at risk.

One particular set of tools that we aim to enable this dataset is the broad creation of multi-frame super-resolution algorithms, to amplify the use of the freely accessible but low-resolution satellite imagery from the European-funded Sentinel 2 constellation. We achieve this by pairing high-resolution from Airbus SPOT 6/7 satellites with temporally-matched low-resolution imagery from Sentinel 2.

To further make machine learning on satellite imagery accessible, we accompany this dataset with an open-source Python package to: rebuild or extend the WorldStrat dataset, train and infer baseline algorithms, and learn with abundant tutorials, all compatible with the popular EO-learn toolbox. Our code for deep learning algorithms provided as baseline on this dataset trains in 60 minutes on a single V100 GPU.

We therefore hope to foster broad-spectrum applications of ML to satellite imagery, and possibly develop from free public low-resolution Sentinel2 imagery the same power of analysis allowed by costly private high-resolution imagery. We illustrate this specific point by training and releasing several highly compute-efficient baselines on the task of Multi-Frame Super-Resolution.

See The Introduction in the main body of the article for a longer perspective – we do not copy it here to avoid redundancy.

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

This dataset was created by:

- Julien Cornebise, Ph.D., Honorary Associate Professor at University College London, and CEO of Why How Ltd, his sole-owner scientific consulting company.
- Ivan Oršolić, independent contractor, working for Why How Ltd for this project.
- Freddie Kalaitzis, Ph.D., Senior Research Fellow at Oxford University, working for Why How Ltd for this project.

We were empowered by the European Space Agency's Phi-Lab, https://philab.phi.esa.int/.

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 this dataset was funded by the European Space Agency (ESA), as part of the "QueryPlanet" project 4000124792/18/I-BG CCN3, with ESA Phi-Lab championing this project. ESA Third Party Mission (TPM) funded the license extension from Airbus required

for distribution of the SPOT (high-resolution) imagery. Sinergise Ltd contributed in kind by giving free access to a SentinelHub account to facilitate access to the Sentinel Imagery. Julien Cornebise contributed in kind by volunteering his time, and funded part of the costs via his company Why How Ltd.

Any other comments?

Some of the locations listed in this dataset were kindly provided by:

- Jamon Van De Hoek, Ph.D., Associate Professor at Oregon State University who indicated the UNHCR Persons of Concerns dataset.
- Micah Farfour, Special Advisor in Remote Sensing at Amnesty International who provided 22 locations of human rights interest.
- Moritz Besser, Machine Learning Consultant at dida Machine Learning who provided 40 locations of artisanal mining.

Providing locations does not engage their responsibility or that of their employers.

The creation of this dataset transforms, or includes data from pre-existing datasets, in particular:

- Randomly samples a subset of locations from UNHCR People of Concerns [UNHCR, 2021].
- Filters world-wide randomly sampled locations using data from ESA CCI LC [ESA, 2017].
- Filters world-wide randomly sampled locations using data from GHSL SMOD [Florczyk et al., 2019].
- The high-resolution imagery included in this dataset comes from Airbus as part of their SPOT 6/7 product [Airbus, 2013].
- The low-resolution imagery included in this dataset comes from the European Space Agency as part of the Copernicus Sentinel2 product [Drusch et al., 2012].

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.

Each instance represents a patch of land on Earth of 2.5 km^2, i.e. 1581 m per side.

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

There are 3,449 instances.

There are two types with regard to size: 2.5km^2 and 22.5km^2 instances:

- There are 3,388 $\times$ 2.5km^2 instances.
- There are 61 $\times$ 22.5km^2 instances.
- Their combined total is 9,820.57km^2.

The 22.5km^2 instances can be split into a grid of 3-by-3 2.5km^2 instances, which brings the total of 2.5km^2 instances to 3,937.

There are four types with regards to their location source:

- 22 $\times$ 22.5km^2 Amnesty instances or 198 $\times$ 2.5km^2 instances.
- 39 $\times$ 22.5km^2 ASMSpotter instances or 351 $\times$ 2.5km^2 instances.
- 981 $\times$ 2.5km^2 UNHCR instances.
- 2,407 $\times$ 2.5km^2 randomly sampled/stratified instances.

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 population of all possible instances would be the full surface of the Earth, at all possible times – so this dataset is very much a sample.

We detailed our sampling procedure in Section 2 of the accompanying article Cornebise et al. [2022a], with parts duplicated in the rest of this answer for the readers' convenience, including the summary in Figure 1.

We use the first half of the dataset to attempt a systematic, stratified coverage of the world. The question becomes: how do we chose these locations to ensure a "best" application-agnostic dataset for super-resolution?

Sixty percent was taken from the "Settlement" class from the ESA CCI LandCover Product, which we then stratified according to the Global Human Settlement Layer SMOD for different types of urban density, and with marginal distribution proportional to the cubic root of the actual distribution -- to keep the order of classes but diminish the overall imbalance.

Forty percent was taken from all the other IPCC classes, i.e. non-settlement, stratified according to (non-settlement) IPCC class, marginal distribution proportional to the cubic root of the actual distribution, and within each (non-settlement) IPCC class, again stratifying, according to the LCCS class (thinner vegetation typology), again with cubic root proportions.

The second half of the dataset is obtained by sourcing 3,895 sq km around 1,062 Points Of Interest (POIs) from specialists of use-cases ignored by most existing datasets. For the rarer type of POIs, we sample 9 actual images in a non-overlapping grid centered on the POI.

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

Each instance comprises of: date and time, geographical coordinates, high resolution imagery, and multiple low resolution imagery, with specifics as follows.

Date and time at which each satellite image was captured. **Geographical coordinates** of the patch of land, as latitude and longitude coordinates of the center of the image, and as those of the bounding box.

High Resolution imagery: 1 image of a visit at high resolution, captured by Airbus SPOT 6/7 satellites in R. We provide both the orthorectified imagery as preprocessed by SentinelHub[3], and the raw cropped product from Airbus. The latter has black bands on the boundaries due to the lack of orthorectification. Each image has 5 channels: RGB (6 m/pixel), Near Infrared (6 m/pixel), and Pan-chromatic channels (1.5m / pixel), at 1054x1054 pixels at that highest resolution. The date of the visit has been picked at random between 2017 and 2019 amongst the visits whose whole-scene cloud-cover is lower than 5%. Because our AOIs are much smaller than a full SPOT scene, it is not absolutely guaranteed that the actual image has precisely 5% cloud – it is likely to be entirely empty of clouds. This provides a good target image to reconstruct in the case of super-resolution.

We provide two types of preprocessing high-resolution imagery: the "raw" imagery as provided by Airbus, and the orthorectified imagery as preprocessed by SentinelHub.

Low Resolution imagery: 16 Low-Resolution images from distinct revisits by Copernicus Sentinel 2, temporally matched to the High-Resolution image – within 5 days for the temporally closest. All 12 spectral bands are covered, at up to 10 m/pixel. We chose to not filter the low resolution Sentinel 2 revisits by their cloud coverage. This is to try and ensure the training distribution on the low resolution is similar to the real world use cases, where the user will want

[3] https://docs.sentinel-hub.com/api/latest/data/airbus/spot

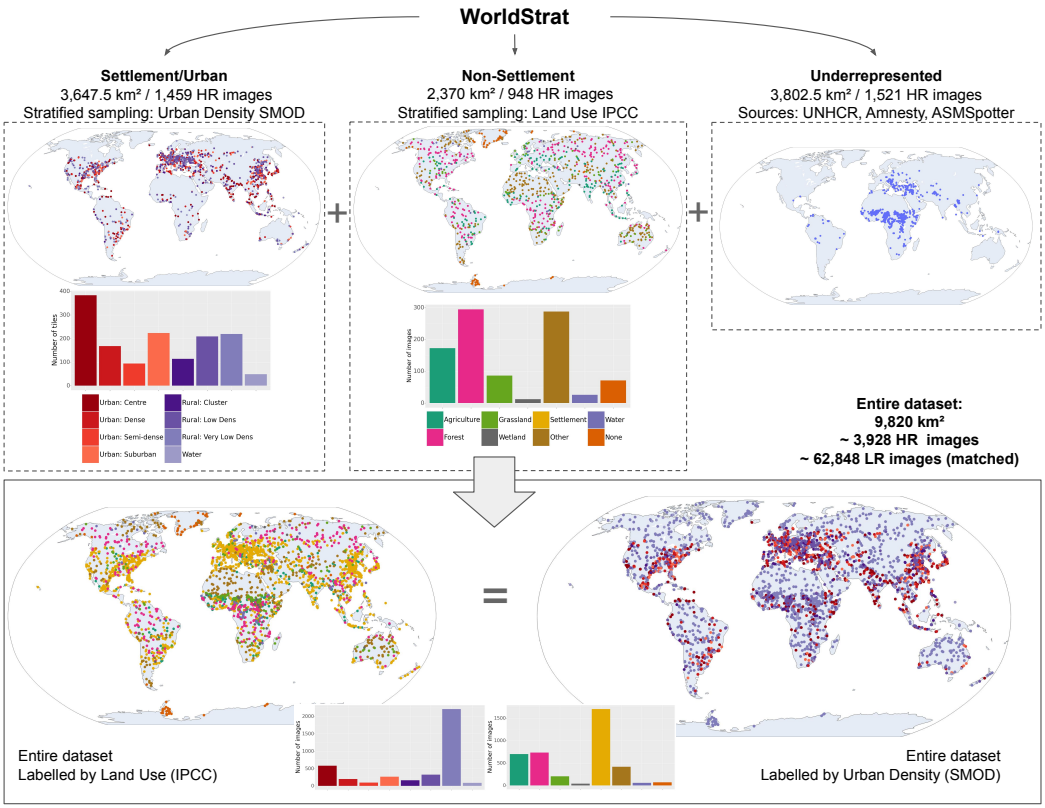

Figure 1: Summarizing the construction and classes of the WorldStrat dataset.

to rebuild at a given place at a given time. Algorithms should learn to ignore clouds and be able to assemble a view from the cloudless parts of the cloudy revisits.

We provide two types of preprocessing for the low-resolution imagery: the orthorectified but non-atmospherically-corrected level "L1C" (according to [ESA, 2015], and level L2A which has been atmospherically corrected, i.e. the effect of the atmosphere on light has been removed according to physical models so colors look closer to the ground conditions.

See below in the Section "Collection" later in this datasheet for **important information about the temporal matching between low and high-resolution imagery**.

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

The low-resolution imagery can be seen as a label for the high-resolution imagery, at least for multi-frame super-resolution tasks. As to formal classes, we provide three labels for each image, coming from the two datasets used to stratify the sampling (see earlier question about sampling):

- **Land use labels** from the European Space Agency (ESA) Climate Change Initiative (CCI) Land Cover (LC) dataset [ESA, 2017], in two forms detailed in Table 2:
  - The highly detailed LCCS classification, with frequencies listed in Table 4 and visualized in Figure 2 (bottom).
  - The matching but coarser Intergovernmental Panel on Climate Change (IPCC) land categories, with frequencies listed in Table 5 (right) and visualized in Figure 2 (top right).

  For more details see page 23 of ESA [2014] and page 32 of ESA [2017].
- **Urban density label** from The Global Human Settlement Layer (GHSL) Settlement Model (SMOD) dataset, indicating the urban density, described in Table 3. Class frequencies are listed in Table 5 (left) and visualized in Figure 2 (top left). For more details see page 24 of Florczyk et al. [2019].

Important Note: The labels correspond to the class assigned to the location at **center of the image** by the corresponding datasets. This *does not mean* that the label is valid for all pixels in the image, because the spatial grids used for each of the two labeling datasets and for the images differ.

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.

Optical satellite observation, always risks suffering from obstruction due to cloud coverage. We have selected the high-resolution images for the lowest cloud-covering at the "scene" level, i.e. the larger product area containing which the 2.5km^2 tile. This lowers considerably the probability of cloud on the small tile we purchased, but does not exclude it entirely. Of course this biases the visit dates towards sunny seasons. We did not do any such filtering on the lower-resolution imaggery, to represent typical sampling conditions.

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.

Any spatial relationship between the instances is given by the geolocation data provided with each tile. Note that, to the best of our verifications, there should be no overlaps between tiles.

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.

IPCC Class	LCCS Class	Class ID
None	No Data	0
Agriculture	Cropland, rain-fed	10
	Herbaceous cover	11
	Tree or shrub cover	12
	Cropland, irrigated or post-flooding	20
	Mosaic cropland (>50%) / natural vegetation (tree, shrub, herbaceous cover) (<50%)	30
	Mosaic natural vegetation (tree, shrub, herbaceous cover) (>50%) / cropland (<50%)	40
Forest	Tree cover, broad-leaved, evergreen, closed to open (>15%)	50
	Tree cover, broad-leaved, deciduous, closed to open (>15%)	60
	Tree cover, broad-leaved, deciduous, closed (>40%)	61
	Tree cover, broad-leaved, deciduous, open (15-40%)	62
	Tree cover, needleleaved, evergreen, closed to open (>15%)	70
	Tree cover, needleleaved, evergreen, closed (>40%)	71
	Tree cover, needleleaved, evergreen, open (15-40%)	72
	Tree cover, needleleaved, deciduous, closed to open (>15%)	80
	Tree cover, needleleaved, deciduous, closed (>40%)	81
	Tree cover, needleleaved, deciduous, open (15-40%)	82
	Tree cover, mixed leaf type (broad-leaved and needleleaved)	90
	Mosaic tree and shrub (>50%) / herbaceous cover (<50%)	100
	Tree cover, flooded, fresh or brackish water	160
	Tree cover, flooded, saline water	170
Grassland	Mosaic herbaceous cover (>50%) / tree and shrub (<50%)	110
	Grassland	130
Wetland	Shrub or herbaceous cover, flooded, fresh/saline/brackish water	180
Settlement	Urban areas	190
Other: Shrubland	Shrubland	120
	Evergreen shrubland	121
	Deciduous shrubland	122
Other: Sparse vegetation	Lichens and mosses	140
	Sparse vegetation (tree, shrub, herbaceous cover) (<15%)	150
	Sparse shrub (<15%)	152
	Sparse herbaceous cover (<15%)	153
Other: Bare area	Bare areas	200
	Consolidated bare areas	201
	Unconsolidated bare areas	202
Other: Water	Water bodies	210
None	Permanent snow and ice	220

Table 2: Land use classes according to the LCSS and the IPCC classifications. LCCS comprises of 23 classes and 14 sub-classes. IPCC groups those into 6 coarser classes.

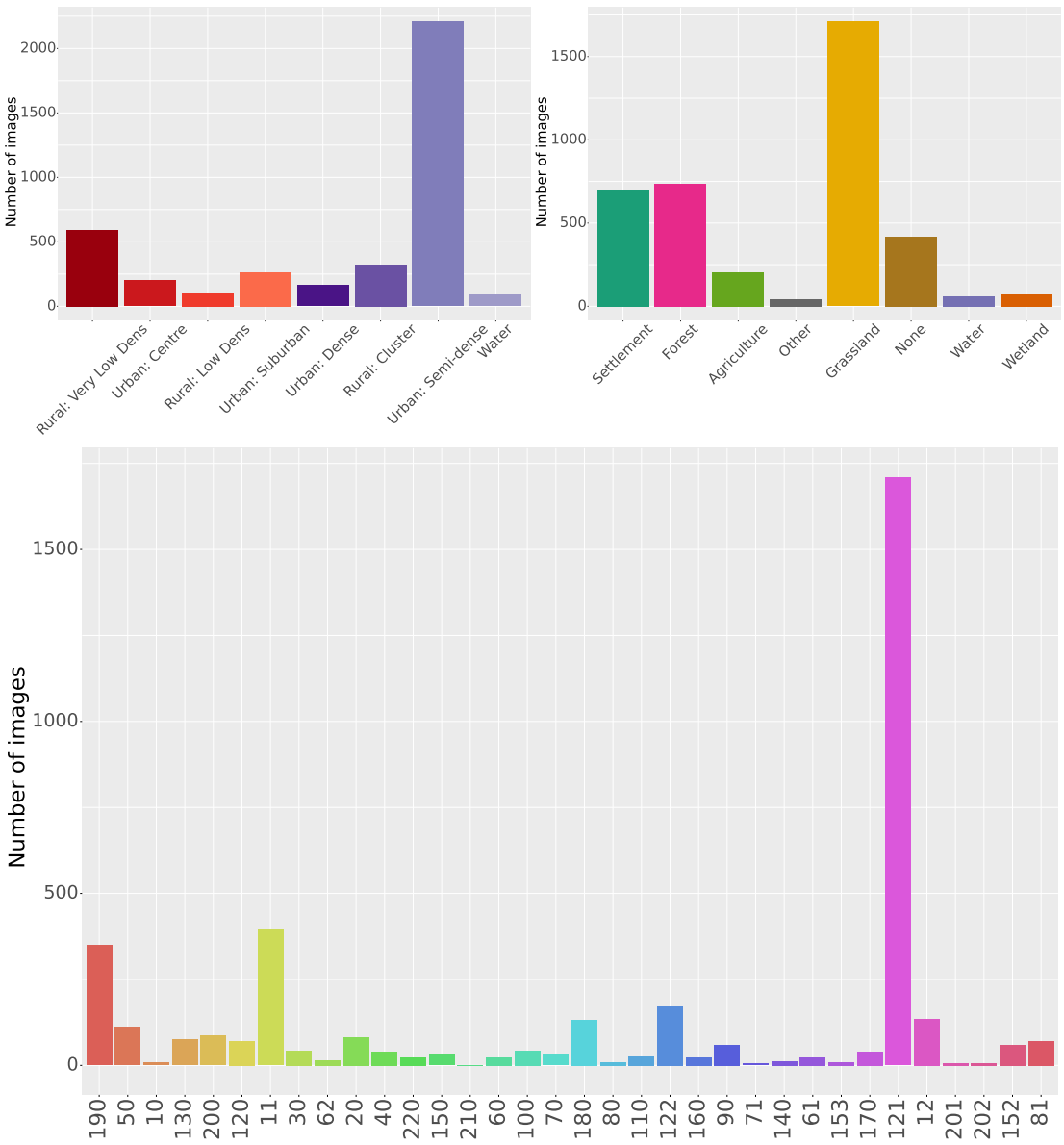

Figure 2: Frequency of class occurrences within the dataset, for SMOD urban density (top left), IPCC land use (top right), LCCS land use (bottom).

SMOD Class	Class ID
30	Urban: Centre
23	Urban: Dense
22	Urban: Semi-dense
21	Urban: Suburban
13	Rural: cluster
12	Rural: Low Dens
11	Rural: Very Low Dens
10	Water

Table 3: The GHSL-SMOD dataset comprises of 3 classes at level 1, and 8 sub-classes at level 2, as described in the GHSL Data Package.

We do provide a recommended split, between training, validation, and testing, for easy comparison. We have used a 80% / 10 % / 10 % proportions amongst the splits, by uniform random sampling amongst all instances.

Note: we have not used stratified sampling to ensure equal class distribution between the three splits. We could and probably should have, and this might get modified in a future version.

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

For the under-represented areas, UNHCR locations are not guaranteed to be locately absolutely precisely on settlements of persons at risk. This is somewhat mitigated by the area of our tiles. In addition, in case they are not, it is important to know that such settlements are often similar to neighbouring constructions: the visual features on the tile can therefore be considered as representative.

As discussed in the previous questions, the spatial grids used for each of the two labeling datasets and for the images have different origins and different resolutions: therefore, all we can guarantee is that the label applies to the corresponding dataset's tile containing central pixel of the imagery.

As also explained in the question on pre-processing, there is a redundancy between the orthorectified and the non-orthorectified high-resolution imagery, and between the L1C and L2A (atmospherically corrected) low-resolution imagery.

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 self-contained.

For ease of use, however, we have included the class information from the ESA CCI LC dataset[ESA, 2017], the UNHCR Persons of Concern dataset [UNHCR, 2021], the images from Copernicus Sentinel 2 archive [Drusch et al., 2012], along with the Airbus SPOT[Airbus, 2013] imagery that we have acquired. None of those have a link to timestamped versions as far as we are aware, hence the inclusion of the subset used in this dataset is more than mere convenience, it is also archiving for consistency.

LCCS Class	Frequency
Urban areas	1,709
Tree cover, broad-leaved, evergreen, closed to open (>15%)	397
Cropland, rain-fed	349
Grassland	172
Bare areas	135
Shrubland	132
Herbaceous cover	111
Mosaic cropland (>50%) / natural vegetation (tree, shrub, herbaceous cover) (<50%)	86
Tree cover, broad-leaved, deciduous, open (15-40%)	82
Cropland, irrigated or post-flooding	75
Mosaic natural vegetation (tree, shrub, herbaceous cover) (>50%) / cropland (<50%)	71
Permanent snow and ice	71
Sparse vegetation (tree, shrub, herbaceous cover) (<15%)	60
Water bodies	59
Tree cover, broad-leaved, deciduous, closed to open (>15%)	43
Mosaic tree and shrub (>50%) / herbaceous cover (<50%)	42
Tree cover, needleleaved, evergreen, closed to open (>15%)	41
Shrub or herbaceous cover, flooded, fresh/saline/brakish water	41
Tree cover, needleleaved, deciduous, closed to open (>15%)	33
Mosaic herbaceous cover (>50%) / tree and shrub (<50%)	33
Deciduous shrubland	29
Tree cover, flooded, fresh or brackish water	24
Tree cover, mixed leaf type (broad-leaved and needleleaved)	23
Tree cover, needleleaved, evergreen, closed (>40%)	23
Lichens and mosses	23
Tree cover, broad-leaved, deciduous, closed (>40%)	15
Sparse herbaceous cover (<15%)	13
Tree cover, flooded, saline water	10
Evergreen shrubland	9
Tree or shrub cover	8
Consolidated bare areas	6
Unconsolidated bare areas	6
Sparse shrub (<15%)	5
Tree cover, needleleaved, deciduous, closed (>40%)	1

Table 4: Frequency of LCCS class occurrences in the dataset.

SMOD Class	Frequency
Rural: Very Low Dens	2,207
Urban: Centre	591
Rural: Low Dens	323
Urban: Suburban	265
Urban: Dense	200
Rural: Cluster	164
Urban: Semi-dense	98
Water	89

IPCC Class	Frequency
Settlement	1,709
Forest	734
Agriculture	700
Other	418
Grassland	205
None	71
Water	59
Wetland	41

Table 5: Frequency of SMOD (left) and IPCC (right) class occurrences in the dataset.

The license on the dataset has been carefully chosen to be compatible with the licenses of each of these original sources: we release the labels and the low-resolution imagery under Creative Commons with Attribution 4.0 International (CC BY 4.0[4], and the high-resolution imagery from Airbus is distributed, with authorization from Airbus, under Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0)[5].

We have used SentinelHub (https://www.sentinel-hub.com/) to download the satellite imagery used in this dataset. All the code used for the construction of the dataset is included in the accompanying Python package.

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.

There is no confidential data in this dataset.

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

This dataset contains satellite imagery of sites classified by UNHCR as hosting Persons of Concern: vulnerable populations such as displaced populations, refugee camps, locations close to conflict zones. It also contains sites of interest for Human Rights investigations, such as prisons or jails.

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

Yes, as there is human density information and inclusion of locations hosting Persons of Concern, but only in aggregate. These locations were already public in the UNHCR Persons of Concerns dataset.

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.

The dataset contains locations and satellite imagery of locations hosting "Persons of Concerns" as classified by UNHCR. This represents 981km^2 out of a total of 9,820km^2 .

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.

No.

[4] https://creativecommons.org/licenses/by/4.0/
[5] https://creativecommons.org/licenses/by-nc/4.0/

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

The location of refugee populations such as listed in UNHCR "Persons of Concerns" dataset and included in this dataset might be considered sensitive. UNHCR already published these locations, and is better placed than us to decide on whether this was sensitive. The locations of some sites of interest for Human Rights investigation, however, are newly listed. And of course, this is the first time that satellite imagery of all these locations is published in one single dataset, to the best of our knowledge – although these are also accessible on common mapping websites for the casual viewer.

See question on ethic reviews later in this datasheet for more details.

Any other comments?

None.

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.

As described above, the imagery was acquired using Airbus SPOT 6/7 and Copernicus Sentinel 2 satellites, downloaded with SentinelHub. We refer to their user guides [Airbus, 2013], as well as to SentinelHub FAQ[6] for details of their internal processing.

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?

We open-source software for the collection of the data in the accompanying Python package, done using parsing of the CSV of the land uses datasets, sampling of locations using pseudo-random number generators and stratified sampling, and SentinelHub API for ordering and downloading the satellite imagery.

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

We refer to the "Composition" section of this datasheet for the description of our sampling methodology.

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 data was assembled by the persons cited as the author of this dataset in the first question. The actual collection of the satellite imagery was done by Airbus and Sentinel 2 / ESA, and we refer to the documentation of the UNHCR, ESA CCI, ASMSpotter, and GHSL SMOD for their respective collection methodologies. The Airbus imagery was acquired as part of ESA Third Party Mission program (TPM) and the license extension for distribution was negotiated and paid for specially by ESA TPM.

[6] https://www.sentinel-hub.com/faq/

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.

- Actual assembly of the dataset took place from March 2021 to June 2022.
- The satellite imagery was filtered to have been taken between 2017 and 2021.
- Each satellite image comes with the timestamp of its acquisition.
- We could not find temporal information about the date of collection by UNHCR of its dataset of locations of "Persons of Concerns". We downloaded thesed locations in September 2021.
- We used the 2020 ESA CCI LC map as base level.
- We used the R2019A release of the GHSL SMOD data [Pesaresi et al., 2019] – see its Product User Guide [Florczyk et al., 2019], sections "Input Data", page 25.

Matching Low-Resolution and High-Resolution Imagery:

Unlike the low-resolution Sentinel2, the high-resolution SPOT is only available where and when it has been tasked. This raises the question of how what date to we pick for the SPOT high-resolution imagery at a location, if multiple are available, and how do we temporally match the Sentinel2 imagery, within which time window around the date of the SPOT visit?

In theory we could try to pick POIs and SPOT visit times that maximise the number of Sentinel2 imagery available within a fixed length time window, so as to have the richest training set. We do indeed observe some variation in that regard: some of these lines go much higher than others.

However, this would induce an implicit bias of a nature hard to interpret. We also observe that within a POI, the discrepancy between the number of S2 revisits, while clearly present, is reasonable, with multiple SPOT revisits offering similar S2 availability.

Finally, biasing per Sentinel2 imagery would be akin to biasing per Sentinel2 cloud coverage: this would not be a fair representation of real-world use cases, and we would therefore be training our models for the wrong problem.

We therefore took the decision yet again to solve a harder problem than an optimally-curated dataset would make for, so as to be the closest to reality. To that effect, within a POI, we pick uniformly at random the SPOT visit to use as a reference.

Of course, one bias remains: we will not have imagery of POIs that have never been tasked by SPOT customers. While unfortunate, there is no way around it, short of using another high resolution product. We do have hope in two mitigating factors:

- SPOT swath covers more than just the single POI, so we cover areas that are possibly more diverse than just the one precise point of interest to the SPOT customer.
- SPOT tasking means the POI exhibits features of activity interesting to at least the SPOT customer. SPOT customers might not have entirely the same interests as the users of our open-source package, but it is not unreasonable to assume that the features will be transferable. Therefore, this implicit sampling is actually a positive way to ensure interesting features.

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 IRB was required. However, we wanted to be careful due to the sensitivity of UNHCR Persons of Concerns locations and the locations provided by Amnesty International. It is important to note that the UNHCR dataset of locations was already published – but not with satellite imagery collated this way.

We therefore consulted with other human rights experts not otherwise involved in this project before releasing this dataset. They pointed at the precedent of, for example, Human Rights Watch (HRW) releasing the map of torture sites in Syria in 2012[7,8]. In that particular case, emphasis was put by the persons on the ground, living near those locations, that these sites were already widely known to local forces. Publication of these locations therefore provided limited extra risk by local exposure, and provided large benefits from the worldwide attention attracted by their global exposure.

Based on these precedents, and, most importantly, on the fact that all this information was already available albeit not in this packaged form, we decided to publish.

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

Only to people in aggregate populations, via SMOD urban density and UNHCR locations.

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

As explained above, we obtained that data about populations via GHSL SMOD and UNHCR.

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.

We do not have access to this information for GHSL SMOD and UNHCR data products. The construction of GHSL SMOD involved in part administrative data such as census, so answering this question accurately would need tracing each individual census they obtained.

We reinforce that we are probably over-cautious in answering this question, as we take a very broad view of the term "relate to people" in the question. These are population-wide density data, and locations of settlements, already published.

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.

See above.

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).

Not applicable

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.

See above about ethical discussion of impact.

Any other comments?

None.

Preprocessing/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,

[7] https://www.hrw.org/video-photos/interactive/2012/07/02/interactive-map-syrias-torture-centers
[8] https://www.hrw.org/news/2012/07/03/syria-torture-centers-revealed

processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

The preprocessing of the imagery was explained above as part of the collection – in remote sensing, the separation between collection and preprocessing is very arbitrary, as they form a continuous spectrum of more and more refined data products.

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.

As described above, we provide both "raw" and orthorectified Airbus data, and L1C and L2A Sentinel 2 imagery.

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

Yes. We provide in the accompanying Python package the entire code we used to collect, sample, assemble, and pre-process the data.

Any other comments?

The processing of satellite imagery is a very complex topic with a long history and a huge domain expertise required - few people master it end-to-end, certainly not us. We wanted to lower the barrier to entry by providing this dataset, and the accompanying PyTorch DataLoader, in a format most accessible to the Machine Learning community. Remote sensing experts might therefore frown upon the levity with which we (do not) discuss many details of satellite imagery (e.g. angle of incidence, atmospheric collection models, etc). We have provided throughout this datasheet the references for anyone interested in tracing all the processing steps by the providers of the individual data products.

Uses

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

Yes. We have used this dataset for a benchmark comparison of three baseline multi-frame super-resolution algorithm, aiming at super-resolving from 10m/pixel (Sentinel 2) multi-spectral to 3m/pixel obtained by down-sampling the Pan-Sharpened SPOT 6/7 imagery. We refer to Section 4 of the accompanying article [Cornebise et al., 2022a], and to the accompanying Python package which allows full reproduction of that benchmark. These are meant as baseline to illustrate how to use this dataset, and we have entire confidence that they will be beat very soon – we are looking forward to users training their own algorithms!

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.

This dataset will be archived on Zenodo with a DOI, which should in theory allow to search for all papers citing it in any bibliographic database such as Google Scholar. Zenodo also maintains automatically a list of uses and citations. It does not allow to manually add uses that do not cite the DOI e.g. because they were not accompanied by a publication. We do not currently have plans to manually maintain a separate repository of usages, but could be convinced to do so if several users request it.

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

As explained in the Introduction of the accompanying article, we have designed this dataset so it can be used for the broadest range of machine learning applications for satellite imagery.

Of course, one immediate use is to further super-resolution research. We believe efficient super-resolution algorithms, in particular from Sentinel 2, can unlock use cases where high-resolution is not available, either due to cost or limited tasking or simply scale – Sentinel 2

being a remarkable resource re-visiting the world every 5 days, accessible to everyone. Our benchmark

Beyond that, we do not pretend to substitute our imagination for the creativity of our colleagues in the community. With the imagery alone, we can imagine any kind of computer vision tasks involving self-supervised or un-supervised representations on low and high resolutions, transfer tasks from one resolution to the other. We could imagine some classification tasks with the labels, with the caveats mentioned earlier on how these were used as a guideline for a rich sampling and might have temporal mismatch. Because every image is geo-referenced and timestamped, it is also possible to cross-reference it with any other source of label, for example mapping databases like OpenStreetMap, for building imprints, structure detection, etc.

By providing the code we used to create the dataset, we also make it very easy to extend its sampling using the same procedure, to obtain imagery new locations, by anyone having access to different high-resolution imagery – Sentinel 2 low-resolution imagery being accessible to everyone already. Whether it be to redistribute or for their own use is up to the user.

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?

We listed throughout this datasheet (and mentioned again in the last answer) several limitations, in particular the temporal matching of the labels with the imagery. That limitation does not impede the variance-reduction and representativity of the dataset, but it does add noise for e.g. classification tasks.

Other than that, there are no impact on future uses that we can think of – and the ability for the user to easily extend the dataset if they get access to new imagery should help ensure its longevity.

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

None to the best of our knowledge.

Any other comments?

None.

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.

Yes, this dataset is made open access as it was designed for the broadest use. We purposefully chose the least restrictive licenses allowed for this dataset to foster reuse and hopefully upstream contributions. Only the high-resolution imagery has a restriction to non-commercial uses, as a requirement from the imagery provider Airbus.

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

The dataset is distributed via Zenodo, a CERN-backed repository for datasets and code, which also provides a DOI for the dataset and a separate DOI for each updated version. Zenodo also takes care of all meta-data formatting for easy discovery.

The DOI of the dataset is: 10.5281/zenodo.6810792.

The accompanying Python package is distributed on Github, and packaged for the popular Python packaging managers (PyPI, Conda, etc). As part of the package, we also distribute several tutorials in the form of Jupyter notebooks.

When will the dataset be distributed?

The dataset is distributed on Zenodo as of July 13th, 2022.

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.

As discussed above, we have worked hard to ensure the broadest diffusion possible, including on the licensing front:

- The high-resolution Airbus imagery is distributed, with authorization from Airbus, under Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0)[9].
- The labels, Sentinel2 imagery, and trained weights are released under Creative Commons with Attribution 4.0 International (CC BY 4.0[10].
- The source code under 3-Clause BSD license[11].

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.

As explained above, while, thanks to ESA Phi-Lab and ESA Third Party Missions, we secured license from Airbus to distribute the high-resolution imagery, that specific part of the dataset is be used only for non-commercial purposes according to the terms of the CC-BY-NC license.

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.

None.

Any other comments?

None.

Maintenance

Who will be supporting/hosting/maintaining the dataset?

Long-term maintenance of the content of the dataset will be by the authors, like every academic.

In terms of hosting, to ensure the maximum availability and long-term life, the dataset is hosted on Zenodo, which is backed by CERN. This is becoming the gold standard in terms of dataset distribution, and provides more than reasonable availability and redundancy. The underlying Zenodo infrastructure and redundancy measures are documented at Zenodo's About - Infrastructure page[12].

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

[9] https://creativecommons.org/licenses/by-nc/4.0/
[10] https://creativecommons.org/licenses/by/4.0/
[11] https://opensource.org/licenses/BSD-3-Clause
[12] https://about.zenodo.org/infrastructure/

Julien Cornebise can be contacted at his email address at University College London: j.cornebise@ucl.ac.uk. In case of any future change of affiliation, he can also be reached at the stable address julien@cornebise.com.

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

There is no erratum – yet! (But we will ensure that first erratum will also contain a correction to this section)

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)?

We will be uploading any modification of the dataset to Zenodo, which will provide a version-specific DOI along with the root fixed DOI covering the ensemble of versions. We do plan to correct errors that we are made aware of, and we welcome contributions of extra imagery!

Zenodo does not yet seem to have a subscription mechanism to automatically notify subscribers of extra information, therefore we will likely post updates on GitHub or set up a mailing list – but this is still to be determined.

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

No retention limits.

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.

Since the dataset is hosted on Zenodo, and Zenodo supports DOI versioning, all the different versions of the dataset are tracked and hosted. The DOI versioning functions similarly to an incremental update, duplicating only the modified files. More details can be found on Zenodo's FAQ, under DOI versioning. [13]

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.

We welcome contributions by any generous users. They are welcome to contact us (see above) to discuss, and we will verify and validate on a case-by-case basis, as well as publish any extra code that will have been used for the enrichment. As mentioned above, we provide the source code we used to build the dataset, which makes it easy for any would-be contributor to help ensuring similar sampling distribution and formatting. Please do contact us ahead of time, we will be delighted to discuss how to enrich this community resource!

Any other comments?

None.

[13] https://help.zenodo.org/#versioning

References

Airbus. SPOT Imagery User Guide. Technical Report SI/DC/13034-v1.0, Airbus DS, July 2013.

J. Cornebise, I. Oršolić, and F. Kalaitzis. Open High-Resolution Satellite Imagery: The WorldStrat Dataset – With Application to Super-Resolution. *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*, 2022a.

J. Cornebise, I. Oršolić, and F. Kalaitzis. The WorldStrat Dataset: Open High-Resolution Satellite Imagery With Paired Multi-Temporal Low- Resolution. *Zenodo*, July 2022b. doi: 10.5281/zenodo.6810792. URL https://doi.org/10.5281/zenodo.6810792.

M. Drusch, U. Del Bello, S. Carlier, O. Colin, V. Fernandez, F. Gascon, B. Hoersch, C. Isola, P. Laberinti, P. Martimort, et al. Sentinel-2: ESA's optical high-resolution mission for GMES operational services. *Remote sensing of Environment*, 120:25–36, 2012.

ESA. Land Cover CCI Product User Guide Version 2.4. Technical report, European Space Agency, 2014.

ESA. Sentinel-2 user handbook, 2015.

ESA. Land Cover CCI Product User Guide Version 2.0. Technical report, European Space Agency, 2017.

A. Florczyk, C. Corbane, D. Ehrlich, S. Freire, T. Kemper, L. Maffenini, M. Melchiorri, M. Pesaresi, P. Politis, M. Schiavina, F. Sabo, L. Zanchetta, European Commission, and Joint Research Centre. *GHSL Data Package 2019: Public Release GHS P2019.* 2019. ISBN 978-92-76-13186-1.

T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. Iii, and K. Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

M. Pesaresi, A. Florczyk, M. Schiavina, M. Melchiorri, and L. Maffenini. GHS settlement grid, updated and refined regio model 2014 in application to ghs-built r2018a and ghs-pop r2019a, multitemporal (1975-1990-2000-2015) r2019a. *European Commission, Joint Research Centre (JRC)*, 10, 2019.

J. Pineau, P. Vincent-Lamarre, K. Sinha, V. Larivière, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and H. Larochelle. The machine learning reproducibility checklist v2.0. Technical report, 2021a.

J. Pineau, P. Vincent-Lamarre, K. Sinha, V. Larivière, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and H. Larochelle. Improving reproducibility in machine learning research: A report from the NeurIPS 2019 reproducibility program. *Journal of Machine Learning Research*, 22, 2021b.

UNHCR. UNHCR People of Concern Dataset - Refugees Operational Data Portal / Web Services, 2021.