BigEarthNet v2.0

The BigEarthNet v2.0 dataset was constructed by the Remote Sensing Image Analysis (RSiM) Group and the Database Systems and Information Management (DIMA) Group at the Technische Universität Berlin (TU Berlin). This work is supported by the European Research Council under the ERC Starting Grant BigEarth and by the Berlin Institute for the Foundations of Learning and Data (BIFOLD).

BigEarthNet v2.0 is a benchmark dataset consisting of 549,488 pairs of Sentinel-1 and Sentinel-2 image patches. To construct BigEarthNet v2.0 with Sentinel-2 image patches (called as BigEarthNet-S2), 115 Sentinel-2 tiles acquired between June 2017 and May 2018 over 10 countries (Austria, Belgium, Finland, Ireland, Kosovo, Lithuania, Luxembourg, Portugal, Serbia, and Switzerland) of Europe were initially selected. All the tiles were atmospherically corrected by the Sentinel-2 Level 2A product generation and formatting tool (sen2cor v2.11). Then, they were divided into 549,488 image patches. Each image patch was associated with a pixel-level reference map and multiple land-cover class labels (i.e., multi-labels) that were derived from the most recent CORINE Land Cover database of the year 2018 (CLC2018 v2020_u1).

To construct BigEarthNet v2.0 with Sentinel-1 image patches (called as BigEarthNet-S1), 312 Sentinel-1 scenes acquired between June 2017 and May 2018 that jointly cover the area of all original 115 Sentinel-2 tiles with close temporal proximity were selected and processed. BigEarthNet-S1 consists of 549,488 preprocessed Sentinel-1 image patches – one for each Sentinel-2 patch.

The BigEarthNet v2.0 dataset includes several significant improvements compared to the previous 1.0 version. These changes include the application of the latest atmospheric correction tool (sen2cor), which results in higher-quality patches. Additionally, the most recent version of the CLC2018 database was utilized to extract label information, overcoming label noise present in BigEarthNet v1.0. Apart from providing patch-level labels, v2.0 additionally includes pixel-level reference maps, making the dataset suitable for pixel- and scene-based learning tasks. Furthermore, BigEarthNet v2.0 introduces a new geographical-based split assignment algorithm, which significantly reduces spatial correlation among the train, validation, and test sets compared to v1.0. For details, please see:

> K. Clasen, L. Hackel, T. Burgert, G. Sumbul, B. Demir, V. Markl, *"reBEN: Refined BigEarthNet Dataset for Remote Sensing Image Analysis", arXiv preprint arXiv:2407.03653, 2024.*

Description of the BigEarthNet v2.0 Structure

Dataset Root Directory Structure

The BigEarthNet v2.0 dataset root directory structure is as follows:

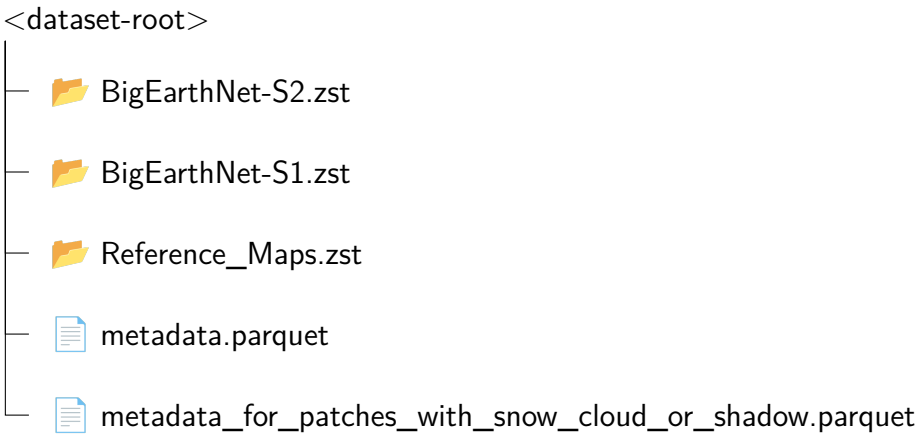

Figure 1: BigEarthNet v2.0 root directory

To minimize storage space and to speed up the transfer of the dataset, each directory of the dataset was compressed with the Zstandard compression algorithm.

BigEarthNet-S2 Directory Structure

The BigEarthNet-S2 directory has one directory per Sentinel-2 source tile (in total 115) and a dedicated directory for each individual patch. Each patch directory contains the GeoTIFF files for the individual bands.

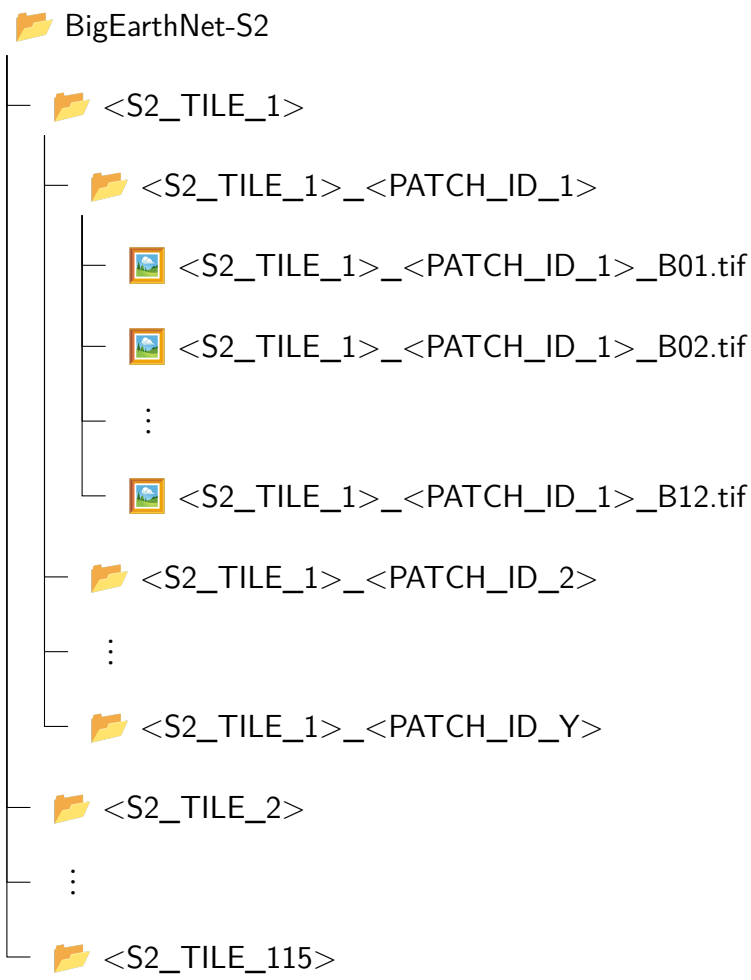

Figure 2: BigEarthNet-S2 directory structure

BigEarthNet-S2 Naming Conventions

The compact naming convention for each BigEarthNet-S2 patch directory is defined as follows:

<Sentinel-ID>_MSIL2A_<YYYYMMDD>T<HHMMSS>_N9999_<Rooo>_↩<Txxxxxx>_<H-Order>_<V-Order>

The components of each folder name are defined as follows:

Sentinel-ID denotes the Sentinel-2 mission ID that can be either `S2A` or `S2B`.

MSIL2A denotes the 2A product level of Sentinel-2 tiles.

YYYYMMDD is the start of the datatake sensing time of a Sentinel-2 tile, including year, month, and day information. For instance, 20170613 denotes "June 13th, 2017".

HHMMSS is the start of the datatake sensing time of a Sentinel-2 tile, including hour, minute and second information. For the time convention, a 24-hour clock format is used. For instance, 140321 denotes "2:03:12 pm".

N9999 is the *processing baseline* and 9999 indicates that sen2cor was executed manually. BigEarthNet v2.0 used sen2cor version 2.11 to generate the L2A tiles.

Roooo is the relative orbit number (R001 - R143).

Txxxxx is the Sentinel-2 tile number field given by ESA.

H-Order identifies the horizontal order of the patch in the tile from which the patch is extracted. This number starts at 0.

V-Order identifies the vertical order of the patch in the tile from which the patch is extracted. This number starts at 0.

Reference Maps Directory Structure

The reference maps directory structure and naming convention are identical to the BigEarthNet-S2 directory structure. The patch directory contains the reference map of the given patch with the `_reference_map.tif` suffix.

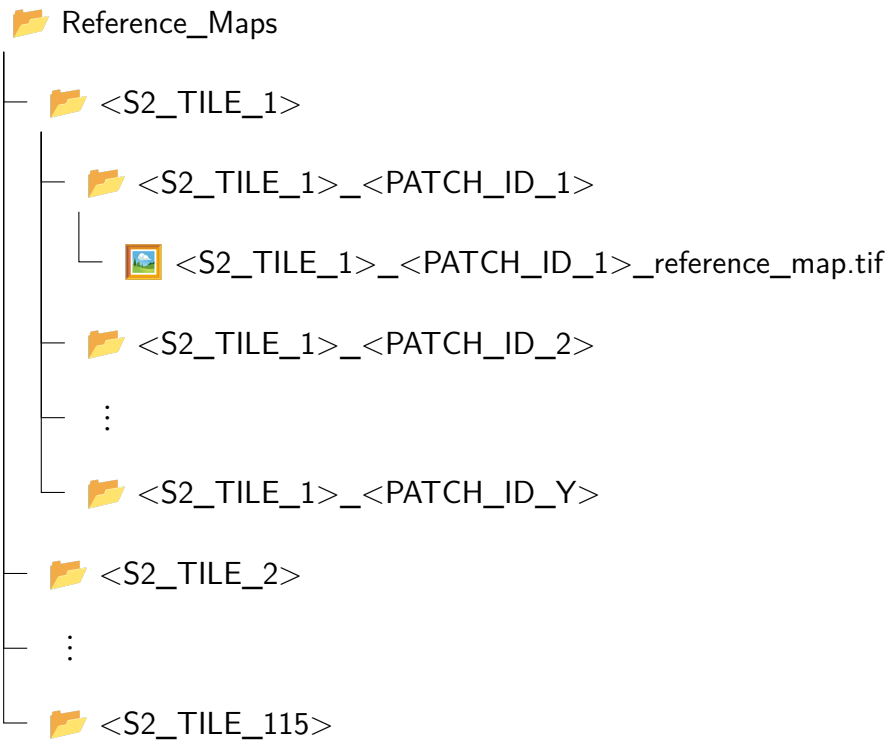

Figure 3: Reference Maps directory structure

Each pixel in the reference map is associated to either one of the CLC Level-3 class labels or with an additional `Unlabeled` class label (which indicates a pixel that does not have an associated CLC label). To derive the recommended 19 class labels for the reference maps, please use the mapping provided in Table 1 together with the class IDs. For more details regarding the 19 classes nomenclature, please see the BigEarthNet-MM publication:

> G. Sumbul, A. d. Wall, T. Kreuziger, F. Marcelino, H. Costa, P. Benevides, M. Caetano, B. Demir, V. Markl, *"BigEarthNet-MM: A Large Scale Multi-Modal Multi-Label Benchmark Archive for Remote Sensing Image Classification and Retrieval", IEEE Geoscience and Remote Sensing Magazine, 2021, doi: 10.1109/MGRS.2021.3089174.*

Table 1: Reference map class IDs associated to the CLC class nomenclature and 19 classes nomenclature

Class ID	CLC Level-3 Class Nomenclature	19 Classes Nomenclature
111	Continuous urban fabric	Urban fabric
112	Discontinuous urban fabric	Urban fabric
121	Industrial or commercial units	Industrial or commercial units
122	Road and rail networks and associated land	Unlabeled
123	Port areas	Unlabeled
124	Airports	Unlabeled
131	Mineral extraction sites	Unlabeled
132	Dump sites	Unlabeled
133	Construction sites	Unlabeled
141	Green urban areas	Unlabeled
142	Sport and leisure facilities	Unlabeled
211	Non-irrigated arable land	Arable land
212	Permanently irrigated land	Arable land
213	Rice fields	Arable land
221	Vineyards	Permanent crops
222	Fruit trees and berry plantations	Permanent crops
223	Olive groves	Permanent crops
231	Pastures	Pastures
241	Annual crops associated with permanent crops	Permanent crops
242	Complex cultivation patterns	Complex cultivation patterns
243	Land principally occupied by agriculture, with significant areas of natural vegetation	Land principally occupied by agriculture, with significant areas of natural vegetation
244	Agro-forestry areas	Agro-forestry areas
311	Broad-leaved forest	Broad-leaved forest
312	Coniferous forest	Coniferous forest
313	Mixed forest	Mixed forest
321	Natural grassland	Natural grassland and sparsely vegetated areas
322	Moors and heathland	Moors, heathland and sclerophyllous vegetation
323	Sclerophyllous vegetation	Moors, heathland and sclerophyllous vegetation
324	Transitional woodland/shrub	Transitional woodland, shrub
331	Beaches, dunes, sands	Beaches, dunes, sands
332	Bare rock	Unlabeled
333	Sparsely vegetated areas	Natural grassland and sparsely vegetated areas
334	Burnt areas	Unlabeled
335	Glaciers and perpetual snow	Unlabeled
411	Inland marshes	Inland wetlands
412	Peatbogs	Inland wetlands
421	Salt marshes	Coastal wetlands
422	Salines	Coastal wetlands
423	Intertidal flats	Unlabeled
511	Water courses	Inland waters
512	Water bodies	Inland waters
521	Coastal lagoons	Marine waters
522	Estuaries	Marine waters
523	Sea and ocean	Marine waters
999	Unlabeled	Unlabeled

BigEarthNet-S1 Directory Structure

The BigEarthNet-S1 directory has one directory per Sentinel-1 source tile (in total 312) and a dedicated directory for each individual patch. Each patch directory contains the GeoTIFF files for the individual bands.

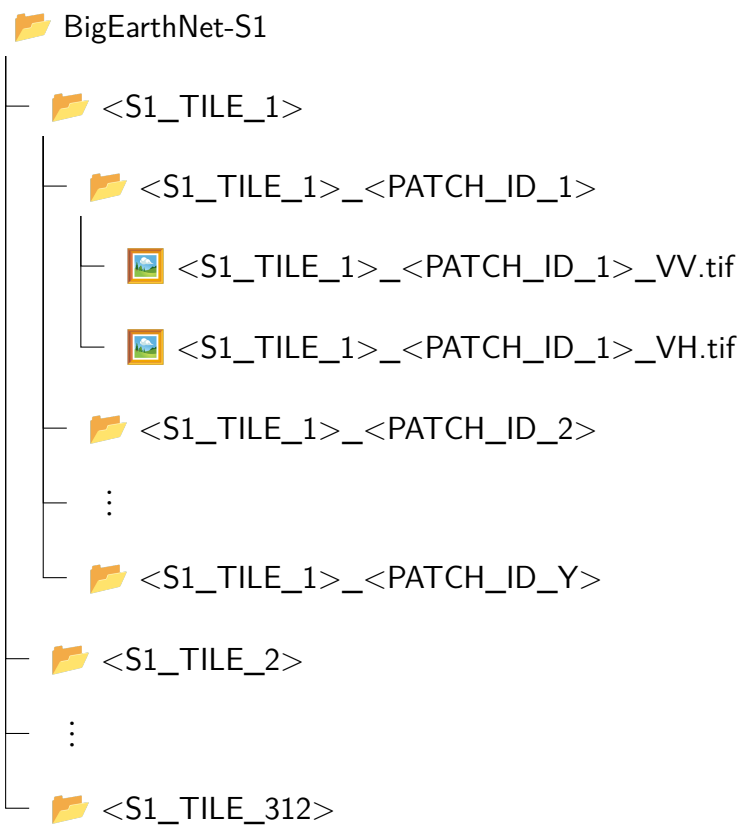

Figure 4: BigEarthNet-S1 Directory Structure

BigEarthNet-S1 Naming Conventions

The compact naming convention for each BigEarthNet-S1 patch directory is defined as follows:

<Sentinel-ID>_IW_GRDH_1SDV_<YYYYMMDD>T<HHMMSS>_↩
<Txxxxxx><H-Order>_<V-Order>

The components of each folder name are defined as follows:

Sentinel-ID denotes the Sentinel-1 mission ID that can be either `S1A` or `S1B`.

IW denotes the acquisition mode of the source product. BigEarthNet-S1 only uses scenes recorded by the *Interferometric Wide Swath* (IW) mode.

GRDH describes the original product type and resolution class. BigEarthNet-S1 only uses *Ground Range Detected* (GRD) in high-resolution (H) images.

1SDV denotes the original processing level (1), product class (S for standard), and polarization (DV for *Dual VV+VH*).

YYYYMMDD is the start of the datatake sensing time of a Sentinel-1 tile, including year, month, and day information. For instance, `20170613` denotes "June 13th, 2017".

HHMMSS is the start of the datatake sensing time of a Sentinel-1 tile, including hour, minute and second information. For the time convention, a 24-hour clock format is used. For instance, 140321 denotes "2:03:12 pm".

Txxxxxx denotes the Sentinel-2 tile id from which the corresponding Sentinel-2 patch has been derived and is given by the ESA.

H-Order identifies the horizontal order of the patch in the tile from which the patch is extracted. This number starts at 0.

V-Order identifies the vertical order of the patch in the tile from which the patch is extracted. This number starts at 0.

File Formats and Metadata Description

For BigEarthNet-S1 and BigEarthNet-S2, each band is stored in a separate GeoTIFF file as a georeferenced raster image. The names are derived by adding band names together with the GeoTIFF extension to the patch folder name.

The parquet[1] files in the BigEarthNet v2.0 root directory provide additional information about the individual patches.

To inspect the parquet files, we recommend using duckdb, polars, or pandas. The columns of the parquet files are described in the following:

patch_id is the BigEarthNet-S2 patch name without the GeoTIFF suffix (example value: `S2B_MSIL2A_20180421T100029_N9999_R122_T33TWM_00_00`).

labels lists the patch-level labels associated with the patch (example value: `[Pastures, Arable land, Mixed forest]`).

split defines to which split (train, validation, test) the patch belongs to (example value: `test`).

country defines to which country a patch is assigned to. The value was derived by inspecting the geographical location of each patch in relation to the administrative country borders from the Natural Earth v5.1.2 110 m dataset (example value: `Austria`).

s1_name defines for each BigEarthNet-S2 patch the associated BigEarthNet-S1 patch name (example value: `S1B_IW_GRDH_1SDV_20170612T165809_33UUP_26_57`).

s2v1_name provides the associated BigEarthNet-S2 v1.0 patch name for a given BigEarthNet-S2 v2.0 patch (example value: `S2A_MSIL2A_20170613T101031_3_51`).

contains_seasonal_snow indicates whether or not the patch contains seasonal snow (example value: `true`).

contains_cloud_or_shadow indicates whether or not the patch contains clouds or cloud shadows (example value `false`).

The **metadata.parquet** file contains the metadata of all of the recommended patches for training and evaluating machine learning models using the BigEarthNet-S2 dataset, which means that the file does not contain metadata about patches that are covered by seasonal snow or by clouds or cloud shadows. The metadata for the remaining patches is stored in the **metadata_for_patches_with_snow_cloud_or_shadow.parquet** file.

[1] Parquet files are an open-source, column-oriented data file format designed for efficient data storage.

Additional Information for the Construction of BigEarthNet-S1

In order to construct the BigEarthNet-S1 patches, 312 Sentinel-1 Ground-Range-Detected (GRD) scenes acquired between June 2017 and May 2018 that jointly cover the area of all original 115 Sentinel-2 tiles with close temporal proximity were selected and processed. All selected scenes are based on the Interferometric Wide (IW) swath mode, which is the main acquisition mode over land.

All scenes were corrected by using the Sentinel-1 Toolbox (S1TBX) and the Graph Processing Framework (GPF) of ESA's Sentinel Application Platform (SNAP). The applied preprocessing workflow includes the application of precise orbit files, border and thermal noise removal, radiometric calibration, and the orthorectification (Range Doppler Terrain Correction) to project the images from slant range to ground range. Based on the spatial extent of the scene, we either employed the SRTM 30 (for scenes below 60° latitude) or the ASTER DEM (for scenes above 60° latitude, where no SRTM 30 exists). Finally, we converted the backscatter coefficient from linear to a decibel (dB) scale for the purpose of data handling. It should be noted that we decided not to apply any speckle filtering in our preprocessing workflow in order to maintain the full resolution. The *Product Family Specification for SAR of the CEOS Analysis Ready Data for Land (CARD4L) framework* also recommends this approach.

Based on the preprocessed Sentinel-1 scenes, for each BigEarthNet-S2 patch, we extracted a corresponding BigEarthNet-S1 patch with the closest possible timestamp. The resulting data patches have a resolution of 10 m x 10 m per pixel and are therefore aligned with the RGB-IR channels of the corresponding BigEarthNet-S2 patches.

At the time of writing, a thorough visual inspection for quality control of the BigEarthNet-S1 patches has not been conducted. Consequently, certain patches may be contaminated by artefacts, such as those caused by interference (also known as Radio-Frequency-Interference RFI) or other dataset-related issues.