

# 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 (DIMIS) 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 Big-EarthNet Dataset for Remote Sensing Image Analysis*”, *arXiv preprint arXiv:2407.03653*, 2024.

The BigEarthNet dataset is licensed under the Community Data License Agreement – Permissive, Version 1.0.

# Description of the BigEarthNet v2.0 Structure

## Dataset Root Directory Structure

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

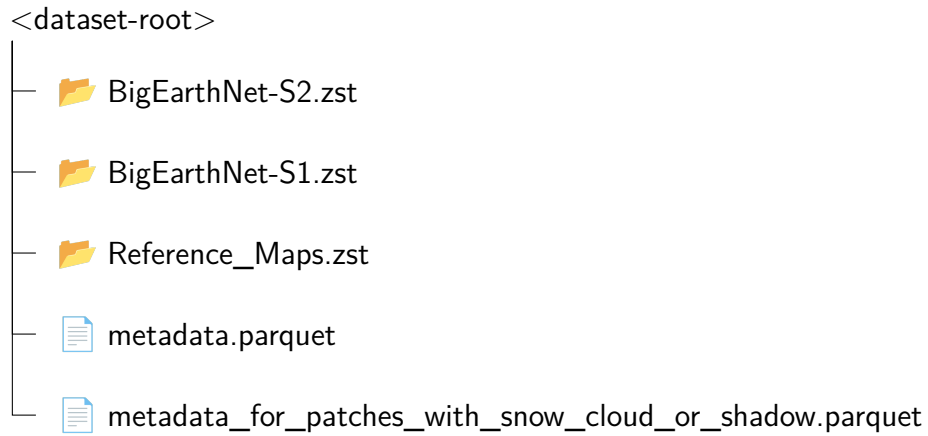


Figure 1: BigEarthNet v2.0 root directory

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.

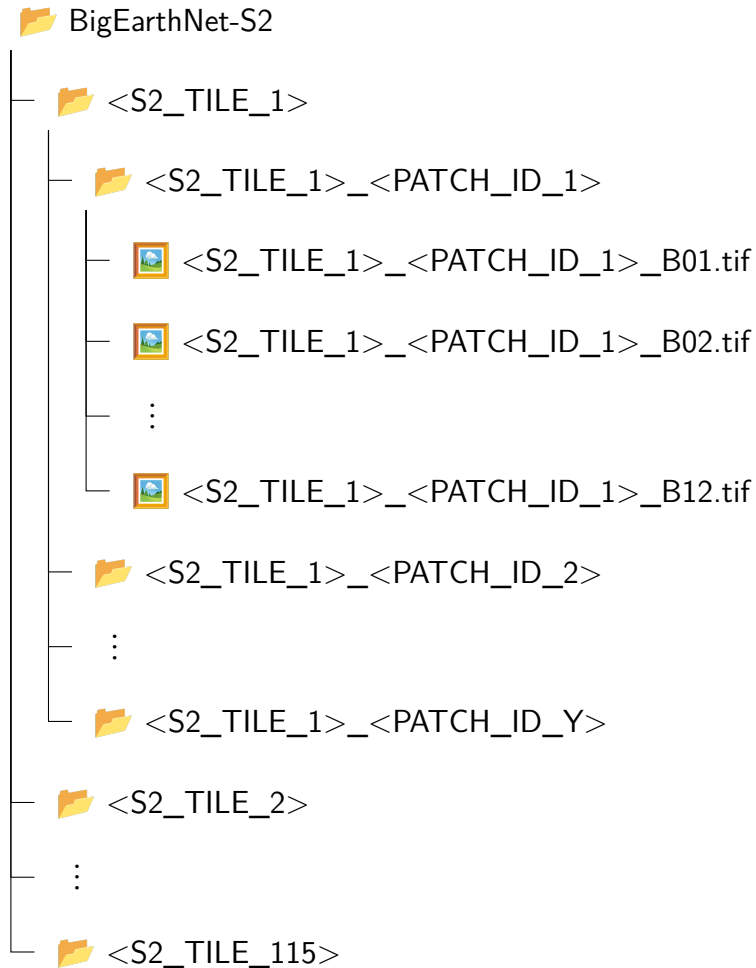


Figure 2: BigEarthNet-S2 directory structure

## 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\_<Roo0>\_<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 6th, 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. BigEarth-Net v2.0 used sen2cor version 2.11 to generate the L2A tiles.

**Ro000** is the relative orbit number (R001 - R143).

**Txxxxxx** 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.

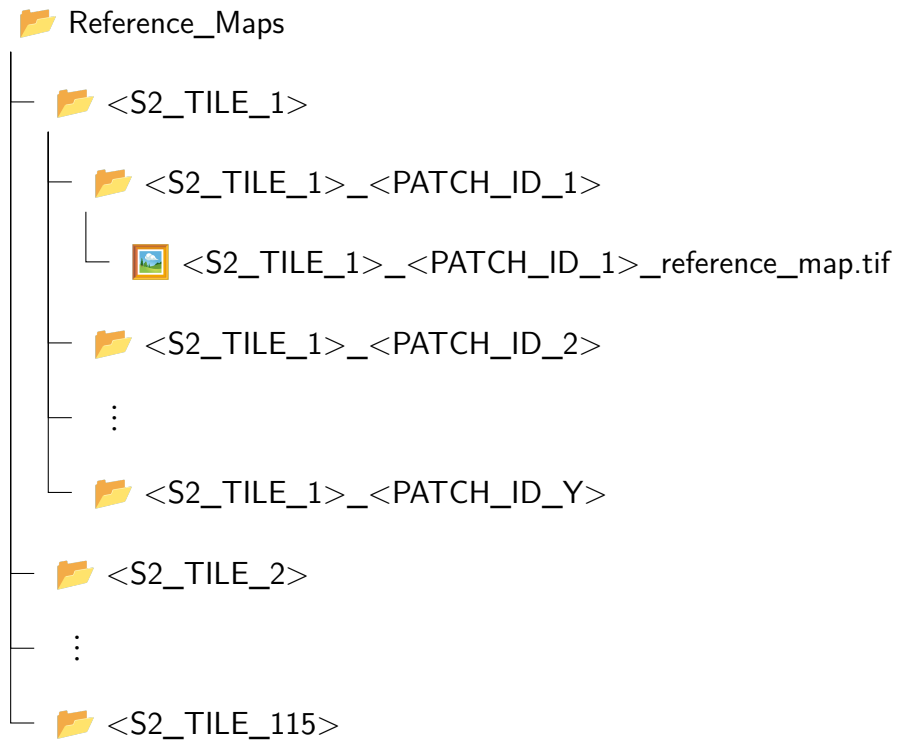


Figure 3: Reference Maps directory structure

## 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.

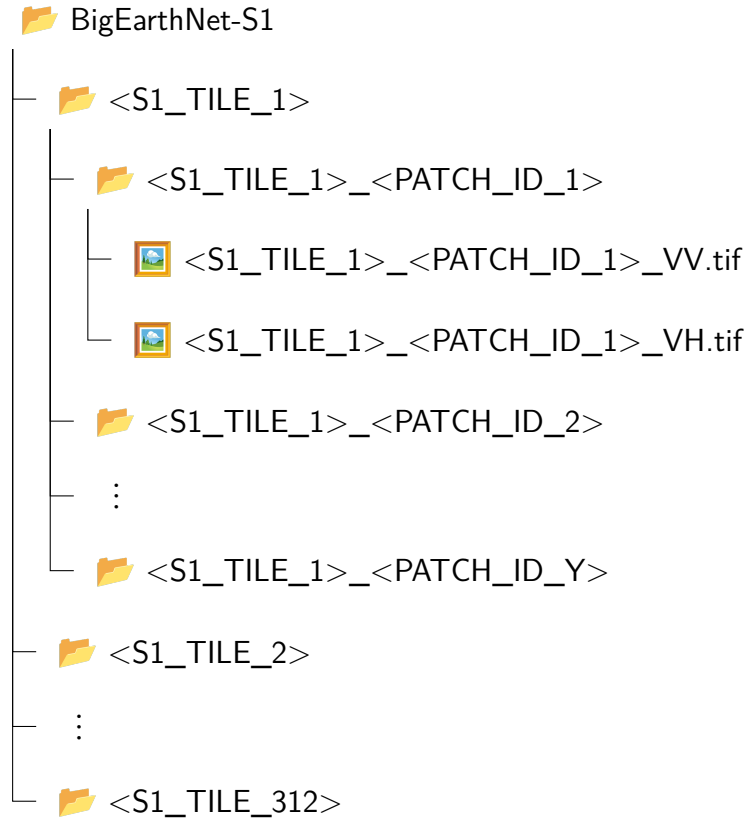


Figure 4: BigEarthNet-S1 Directory Structure

## 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 6th, 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<sup>1</sup> files in the BigEarthNet v2.0 root directory provide additional information about the individual patches.

To inspect the parquet files, we recommend using duckdb, <https://pola.rs/>, 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.

---

<sup>1</sup>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.