The BigEarthNet Archive

The BigEarthNet archive 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). Before BIFOLD, the Berlin Big Data Center (BBDC) supported the work.

BigEarthNet is a benchmark archive, consisting of 590,326 pairs of Sentinel-1 and Sentinel-2 image patches. The first version (v1.0-beta) of BigEarthNet includes only Sentinel 2 images. Recently, it has been enriched by Sentinel-1 images to create a multi-modal BigEarthNet benchmark archive (called also as BigEarthNet-MM).

To construct BigEarthNet with Sentinel-2 image patches (called as BigEarthNet-S2 now, previously BigEarthNet), 125 Sentinel-2 tiles acquired between June 2017 and May 2018 over the 10 countries (Austria, Belgium, Finland, Ireland, Kosovo, Lithuania, Luxembourg, Portugal, Serbia, Switzerland) of Europe were initially selected. All the tiles were atmospherically corrected by the Sentinel-2 Level 2A product generation and formatting tool (sen2cor). Then, they were divided into 590,326 non-overlapping image patches. Each image patch was annotated by the multiple land-cover classes (i.e., multi-labels) that were provided from the CORINE Land Cover database of the year 2018 (CLC 2018).

To construct BigEarthNet with Sentinel-1 image patches (called as BigEarthNet-S1), 321 Sentinel-1 scenes acquired between June 2017 and May 2018 that jointly cover the area of all original 125 Sentinel-2 tiles with close temporal proximity were selected and processed. BigEarthNet-S1 consists of 590,326 preprocessed Sentinel-1 image patches - one for each Sentinel-2 patch. A more detailed explanation on the processing is given in the following pages.

If you use BigEarthNet, please cite:

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.

Description of the BigEarthNet Sentinel-1 Structure

Dataset Information

The BigEarthNet-S1 benchmark archive extends the BigEarthNet-S2 benchmark archive by corresponding Sentinel-1 Synthetic Aperture Radar (SAR) patches to create a multi-modal benchmark archive. In order to construct BigEarthNet-S1, 321 Sentinel-1 Ground-Range-Detected (GRD) scenes acquired between June 2017 and May 2018 that jointly cover the area of all original 125 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 is noted that, since the selection of the speckle filter is considered to be application dependent, no speckle filtering was applied in our preprocessing workflow in order to preserve the full resolution. This approach is also recommended by the *Product Family Specification for SAR* of the *CEOS Analysis Ready Data for Land (CARD4L)* framework[1].

Based on the preprocessed Sentinel-1 scenes, for each Sentinel-2 patch we extracted a corresponding Sentinel-1 patch with the closest possible timestamp. In addition, each Sentinel-1 patch inherited the annotations of the corresponding Sentinel-2 patch (i.e. multi-labels provided by the CLC 2018). The resulting data patches have a resolution of 10x10 meters per pixel and are therefore aligned with the RGB-IR channels of the corresponding Sentinel-2 patches.

To the time of this writing, no complete visual inspection for quality control of the dataset has been performed. Therefore, some data patches can be contaminated by artefacts, e.g., caused by interference, which is also known as Radio-Frequency-Interference (RFI), or other dataset related issues.

Archive Directory Structure

Each image patch in BigEarthNet has one directory under the archive root directory. Each patch directory contains the GeoTIFF files for the VV and VH polarisation data of Sentinel-1 and a JSON file containing the corresponding metadata and labels. The directory structure of the archive is shown in the following table. Different levels of the directory hierarchy are shown in different colors.

[1] https://ceos.org/ard/

Directory Hierarchy	Description
<archive-root>/	<archive-root> path: BigEarthNet-S1/
S1A_IW_GRDH_1SDV_20180327T064301_0_0/	Directory of the patch S1A_IW_GRDH_1SDV_20180327T064301_0_0
S1A_IW_GRDH_1SDV_20180327T064301_0_0_VV.tif	GeoTIFF file of the VV band for the patch S1A_IW_GRDH_1SDV_20180327T064301_0_0
S1A_IW_GRDH_1SDV_20180327T064301_0_0_VH.tif	GeoTIFF file of the VV band for the patch S1A_IW_GRDH_1SDV_20180327T064301_0_0
S1A_IW_GRDH_1SDV_20180327T064301_0_0_labels_metadata.json	JSON file containing labels and metadata information for the patch S1A_IW_GRDH_1SDV_20180327T064301_0_0
S1A_IW_GRDH_1SDV_20180327T064301_0_1/	Directory of the patch S1A_IW_GRDH_1SDV_20180327T064301_0_1
...	
...	

Naming Conventions

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

<sentinel-id>_IW_GRDH_1SDV_<YYYYMMDD>T<HHMMSS>_<utm_tile>
<h-order><v-order>

Components of each folder name are defined as follows:

- **<sentinel-id>** is 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 polarisation (DV for Dual VV+VH).
- **<YYYYMMDD>** is the acquisition date of a Sentinel-2 tile including year, month, and day information. For instance, 20170717 denotes 'July 17th, 2017'.
- **<HHMMSS>** is the acquisition time of a Sentinel-2 tile including hour, minute, and second information. For the time convention, a 24-hour clock format is used.
- **<utm_tile>** denotes the tile identifier of the Sentinel-2 UTM tiling grid from which the corresponding Sentinel-2 patch has been derived.
- **<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

- Each band is stored in a separate GeoTIFF file as a georeferenced raster image. Names of these files are defined by adding the band names together with the .tif extension (_VV.tif, _VH.tif) to the patch folder names.

- Multi-labels and metadata of each patch are stored in a JSON file with a name extension (_labels_metadata.json) to the patch folder name. Name-value pairs and their explanation are as follows:

{

"labels": ["<label_1>", "<label_2>",…,"<label_3>"], (*Multiple class names in an array structure*),

"scene_source": "<S1_GRD_Product_Name>", (*Original unprocessed Sentinel-1 scene, which can be obtained from Copernicus Open Access Hub, product name*),

"acquisition_date": "<YYYY-MM-DDTHH:MM:SS>", (*Acquisition date and time of the corresponding tile*),

"coordinates": **{** (*Upper left and lower right corner coordinates of the patch*)

"ulx" : <upper_left_x>,

"uly" : <upper_left_y>,

"lrx" : <lower_right_x>,

"lry" : <lower_right_y>

},

"projection": "<wkt_projection>" (*Projection for the patch coordinates in Well-Known Text format (WKT)*),

"corresponding_s2_patch": <s2_patch_name> (*Name of the corresponding S2 patch of the original BigEarthNet Archive that covers the exact same area.*)

}