RapidAI4EO

A Corpus of Dense Time Series Satellite Imagery

Document version: 1.0

1. Overview

The RapidAI4EO Corpus is a dataset of dense time series satellite imagery sampled at 500,000 locations across Europe that was produced in the context of the European Union Horizon 2020 project RapidAI4EO. The goal of the project was to establish the foundations for the next generation of land monitoring applications, and as such the corpus was designed to support the development of machine learning models for land use and land cover (LULC) classification and change detection in the ontology of the CORINE land cover (CLC) inventory. The 500,000 locations were chosen to correspond to the area representation of the countries in the European Environmental Agency's EEA39 product, and to maintain the distribution of the 44 CLC classes. The corpus is now being released as open data to support research in the domains of LULC classification and change detection as well as other topics that could benefit from dense time series satellite imagery.The corpus contains the following:

Content	Role
Planet Fusion time series imagery	Core training data
Sentinel-2 time series imagery	Core training data
Planet Fusion quality assurance layers	Planet Fusion metadata
Sentinel-2 traceability files	Sentinel-2 metadata
CLC multiclass labels	Datacube-level annotations for LULC machine learning
Spatiotemporal Asset Catalog (STAC)	Facilitate browsing and access

2. Imagery

The corpus is comprised of temporal datacubes of satellite imagery at 500,000 locations. At each location, there is a monthly time series of Sentinel-2 medium-resolution image mosaics covering the calendar year 2018, and a five-day cadence time series of Planet Fusion high-resolution imagery covering 2018—2019. Each timestep in the datacube is stored as a single GeoTIFF image file in the coordinate reference system of the local UTM zone as defined by the Planet Fusion tiling grid. The datacubes have a spatial footprint of 600 × 600 meters at each location.

Product	Period	Spatial resolution	Spectral resolution	Temporal cadence
Planet Fusion	2018—2019	3m	4 bands	Five-day
Sentinel-2 mosaics	2018	10m	12 bands	Monthly

Planet Fusion is an analysis-ready data product consisting of spectrally harmonized and gap-filled time series imagery. Harmonization is conducted against reference satellite systems, such as Sentinel-2 and Landsat 8, to produce a spectrally-consistent time series. Additionally, Planet Fusion applies a sophisticated gap-filling approach to remove pixels contaminated by clouds and cloud shadows, resulting in a spatially complete and temporally continuous product. Quality assurance layers regarding gap-filling are also produced, as detailed in **Metadata**.

Sentinel-2 monthly image mosaics were produced by ingesting the Sentinel-2 MSI L2A product and taking the pixelwise-temporal median of all observations for each month after applying the sen2cor scene classification layer for cloud masking. The resulting mosaics have the same format as the MSI

L2A product.Band information for both imagery products is included in **Appendix 1**. Additionally, Planet Fusion details are available in the , and the specification of the Sentinel-2 MSI L2A product is available .

3. METADATA

Metadata included in the corpus are comprised of Planet Fusion quality assurance layers and Sentinel-2 traceability files. Planet Fusion quality assurance layers contain pixelwise information relative to gap-filling and harmonization, such as whether a pixel was gap filled and which observations were available to inform the harmonization as detailed in **Appendix 1**. These can be used as masks in any downstream tasks. Sentinel-2 traceability files list all Sentinel-2 MSI L2A scenes that were used to create image mosaics at the level of the Planet Fusion tile. From these tile-level mosaics the final Sentinel-2 datacubes were cropped. Additional metadata, such as the spatial and temporal extent of data products, are natively captured by the STAC format.

4. LABELS

The labels in the corpus are CLC level-3 multiclass labels. Specifically, each sample is annotated with the percentage of each CLC class from the 2018 product that covers that location, represented as a value in the range [0, 1], with the summation of all labels for a sample always equalling one. As the CLC inventory is updated on a six-year basis, the labels are mono-temporal. Higher-level or bespoke aggregations of the CLC level-3 labels can be achieved by simply summing the multiclass labels corresponding to the new classes.

Users are encouraged to combine the imagery in this corpus with other label sets to conduct research in other LULC ontologies or different research topics altogether. Additionally, the imagery can be used without labels to pretrain models to create spatial or spatio-temporal embeddings that can be applied to downstream tasks.

5. STRUCTURE

The contents of the corpus are described by the included SpatioTemporal Asset Catalog (STAC).[1] At the root level, this consists of a Catalog linking one Collection for each product type:

- Planet Fusion imagery and metadata: `rapidai4eo_v1_source_pf`
- Sentinel-2 imagery and metadata: `rapidai4eo_v1_source_s2`
- CLC multiclass labels: `rapidai4eo_v1_labels`

Nested within each product-level Collection are sub-Collections representing logical spatial partitions in the data based on the tiling grid against which Planet Fusion is produced and therefore the corpus sampling was designed.[2] These sub-Collection serve to optimize spatial queries of the corpus.

The first level of sub-Collections split the data by UTM zone, with one Collection for each of the 12 zones across which imagery and labels were sampled. For example, all Planet Fusion imagery in the UTM zone 34N would be nested within the collection `rapidai4eo_v1_source_pf_34N`. All imagery

[1] For details on STAC elements, standards, and best practices refer to the STAC Specification.

[2] The Planet Fusion tiling grid naming conventions are applied to all products, including Sentinel-2 mosaics and CLC labels.

within a UTM zone Collection is in the same coordinate reference system, specifically the coordinate reference system of the local zone.

Beneath the UTM zone Collections, there is one Collection for each 24 × 24-kilometre tile produced from the Planet Fusion processing grid, represented as its offset within the UTM zone by an easting and a northing. For example, all Planet Fusion imagery in UTM zone 34N and the processing tile with offset 24E-241N would be nested in the collection `rapidai4eo_v1_source_pf_34N_24E-241N`. Each of the 1,402 tile Collections in the corpus belongs to only one UTM zone.

Each tile Collection enumerates the STAC Items sampled from that tile as a two-component offset. Building on the previous examples, an Item representing the Planet Fusion datacube at a single location would take the form: `rapidai4eo_v1_source_pf_34N_24E-241N_15_05`. The spatial components of STAC identifiers are:

34N	24E-241N	15_05
UTM zone	Planet Fusion tile offset within the UTM zone	Sample offset within the grid tile

Imagery products, both Planet Fusion and Sentinel-2, as well as labels are spatially consistent within the corpus, meaning that each product exists for all of the 500,000 corpus sampling locations. With this knowledge, the various products can be tied together by leveraging their STAC identifiers. For example, the following three STAC Items correspond to the various products at the same location:

- Planet Fusion: `rapidai4eo_v1_source_pf_34N_24E-241N_15_05`
- Sentinel-2: `rapidai4eo_v1_source_s2_34N_24E-241N_15_05`
- CLC multiclass labels: `rapidai4eo_v1_labels_34N_24E-241N_15_05`

The data and metadata files themselves are linked to the STAC as Assets, usually at the Item level but with selected Assets at the level of Collections. Imagery, both Planet Fusion and Sentinel-2, is stored as one GeoTIFF file per location and timestep. Therefore, the STAC Items for the imaging products have one image Asset per timestep. Planet Fusion also has one quality assurance file for each timestep, which are linked to the Item as metadata Assets. Labels are stored as GeoJSON features with the multiclass labels given as properties. As corpus labels are monotemporal, each labels Item links a single Asset.

Sentinel-2 metadata, the traceability files, correspond to all scenes that were used to create mosaics at the tile level, and therefore are linked as Assets in the tile Collections, *e.g.* `rapidai4eo_v1_source_s2_34N_24E-241N`. Several additional key Assets are stored within the product Collections to facilitate understanding and querying of the corpus. The labels Collection, `rapidai4eo_v1_labels`, has an Asset mapping labels between the three levels of the CLC hierarchy, as well as an Asset that list the labels at all 500,000 locations. The two imagery collections, `rapidai4eo_v1_source_pf` and `rapidai4eo_v1_source_s2`, both link an Asset that lists the geometries of all 500,000 sample locations. These latter assets allow for quick searches by labels or geometries without requiring traversal of the entire STAC.

6. LICENSING

The corpus contains modified Copernicus Sentinel data for 2018 processed by Vision Impulse. The Sentinel data were modified by creating monthly image mosaics. Sentinel data are free and open for public use under EU law. For full details of use, refer to the Copernicus Sentinel Data Terms and Conditions. Those terms shall apply accordingly to the modified Copernicus Sentinel data.

The multiclass LULC labels in this corpus were adapted by Vision Impulse from the CORINE Land Cover product of the European Union's Copernicus programme. The CLC product, with funding by the European Union, was adapted using an aggregation process. The source product is available here. For full details of use, refer to the Registration and Licensing Conditions of the European Parliament and of the Council on the European Earth monitoring programme. Those terms shall apply accordingly to the aggregated CORINE Land Cover labels.

7. Citation

Davis, T., Bischke, B., Helber, P., Senaras, C., Rana, A., Wania, A., Van De Kerchove, R., Zanaga, D., De Keersmaecker, W., Lesiv, M., Ranera, F., & Marchisio, G. (2023) "RapidAI4EO: A Corpus of Dense Time Series Satellite Imagery", Version 1.0, Radiant MLHub. [**Date Accessed**] https://doi.org/10.34911/RDNT.GCYDKJ.

In BibTeX format:

```
@dataset{rapidai4eo,
  title = {RapidAI4EO: A Corpus of Dense Time Series Satellite Imagery},
  publisher = {Radiant ML Hub},
  doi = {https://doi.org/10.34911/RDNT.GCYDKJ},
  year = {2023},
  urldate = {<date of access, ISO>},
  author = {
    Davis, Timothy and
    Bischke, Benjamin and
    Helber, Patrick and
    Senaras, Caglar and
    Rana, Akhil and
    Wania, Annett and
    Van De Kerchove, Ruben and
    Zanaga, Daniele and
    De Keersmaecker, Wanda and
    Lesiv, Myroslava and
    Ranera, Franck and
    Marchisio, Giovanni
  },
  version = {1.0},
}
```

8. Funding

The RapidAI4EO project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004356.

9. Acknowledgements

The authors wish to thank Radiant Earth Foundation for hosting this corpus and supporting in packaging the data for release.

10. CONTACT

hello@rapidai4eo.eu

Appendix 1: Band Information

Planet Fusion surface reflectance bands

Layer	Description
Band 1	Blue band (0.45 - 0.51 µm) SR (NBAR)
Band 2	Green band (0.53 - 0.59 µm) SR (NBAR)
Band 3	Red band (0.64 - 0.67 µm) SR (NBAR)
Band 4	NIR band (0.85 - 0.88 µm) SR (NBAR)

Planet Fusion quality assurance bands

Layer	Description
Layer 1	Percentage of synthetic data used to generate pixel
Layer 2	Days to closest scene with an actual observation
Layer 3	PlanetScope cloud and cloud shadow mask
Layer 4	PlanetScope pixel traceability mask
Layer 5	Number of reference scenes used during calibration
Layer 6	Blue band uncertainty estimate
Layer 7	Green band uncertainty estimate
Layer 8	Red band uncertainty estimate
Layer 9	NIR band uncertainty estimate

Sentinel-2 MSI L2A surface reflectance bands

Layer	Description
B02	Blue band
B03	Green band
B04	Red band
B08	NIR band
B05	Red edge band 1
B06	Red edge band 2
B07	Red edge band 3
B8A	Narrow NIR band
B11	SWIR band 1
B12	SWIR band 2
B01	Coastal aerosol band
B09	Water vapour band

Appendix 2: CLC Class Hierarchy

Level 1	Level 2	Level 2
1. Artificial Surfaces	1.1 Urban fabric	1.1.1 Continuous urban fabric
		1.1.2 Discontinuous urban fabric
	1.2 Industrial, commercial and transport units	1.2.1 Industrial or commercial units
		1.2.2 Road and rail networks and associated land
		1.2.3 Port areas
		1.2.4 Airports
	1.3 Mine, dump and construction sites	1.3.1 Mineral extraction sites
		1.3.2 Dump sites
		1.3.3 Construction sites
	1.4 Artificial, non-agricultural vegetated areas	1.4.1 Green urban areas
		1.4.2 Sport and leisure facilities
2. Agricultural areas	2.1 Arable land	2.1.1 Non-irrigated arable land
		2.1.2 Permanently irrigated land
		2.1.3 Rice fields
	2.2 Permanent crops	2.2.1 Vineyards
		2.2.2 Fruit trees and berry plantations
		2.2.3 Olive groves
	2.3 Pastures	2.3.1 Pastures
	2.4 Heterogeneous agricultural areas	2.4.1 Annual crops associated with permanent crops
		2.4.2 Complex cultivation patterns
		2.4.3 Land principally occupied by agriculture, with significant areas of natural vegetation
		2.4.4 Agro-forestry areas
3. Forest and seminatural areas	3.1 Forest	3.1.1 Broad-leaved forest
		3.1.2 Coniferous forest
		3.1.3 Mixed forest
	3.2 Shrub and/or herbaceous vegetation associations	3.2.1 Natural grassland
		3.2.2 Moors and heathland
		3.2.3 Sclerophyllous vegetation
		3.2.4 Transitional woodland/shrub
	3.3 Open spaces with little or no vegetation	3.3.1 Beaches, dunes, sands
		3.3.2 Bare rock
		3.3.3 Sparsely vegetated areas
		3.3.4 Burnt areas
		3.3.5 Glaciers and perpetual snow
4. Wetlands	4.1 Inland wetlands	4.1.1 Inland marshes
		4.1.2 Peatbogs
	4.2 Coastal wetlands	4.2.1 Salt marshes
		4.2.2 Salines
		4.2.3 Intertidal flats
5. Water bodies	5.1 Inland waters	5.1.1 Water courses
		5.1.2 Water bodies
	5.2 Marine waters	5.2.1 Coastal lagoons
		5.2.2 Estuaries
		5.2.3 Sea and ocean

Reproduced from https://land.copernicus.eu/user-corner/technical-library/corine-land-cover-nomenclature-guidelines/html/.