OpenSentinelMap: A Large-Scale Land Use Dataset using OpenStreetMap and Sentinel-2 Imagery

Noah Johnson, Wayne Treible, Daniel Crispell
Vision Systems Inc
{noah.johnson, wayne.treible, daniel.crispell}@visionsystemsinc.com

Abstract

Remote sensing data is plentiful, but downloading, organizing, and transforming large amounts of data into a format readily usable by modern machine learning methods is a challenging and labor-intensive task. We present the OpenSentinelMap dataset, which consists of 137,045 unique 3.7 km^2 spatial cells, each containing multiple multispectral Sentinel-2 images captured over a 4 year time period and a set of corresponding per-pixel semantic labels derived from OpenStreetMap data. The labels are not necessarily mutually exclusive, and contain information about roads, buildings, water, and 12 land-use categories. The spatial cells are selected randomly on a global scale over areas of human activity, without regard to OpenStreetMap data availability or quality, making the dataset ideal for both supervised, semi-supervised, and unsupervised experimentation. To demonstrate the effectiveness of the dataset, we a) train an off-the-shelf convolutional neural network with minimal modification to predict land-use and building and road location from multispectral Sentinel-2 imagery and b) show that the learned embeddings are useful for downstream fine-grained classification tasks without any fine-tuning. The dataset is publicly available at https://visionsystemsinc.github.io/open-sentinel-map/.

1. Introduction

Automated methods for analysing the contents of remote sensing imagery are of critical importance given the abundant and constantly growing volume of available data. Modern machine learning approaches for detecting and segmenting regions of interest in remotely sensed images are powerful but typically require large training sets tailored to the specific task at hand, including the geographic region of interest. The goal of the proposed dataset is to support a variety of machine learning approaches and applications using a single large-scale global dataset with annotations

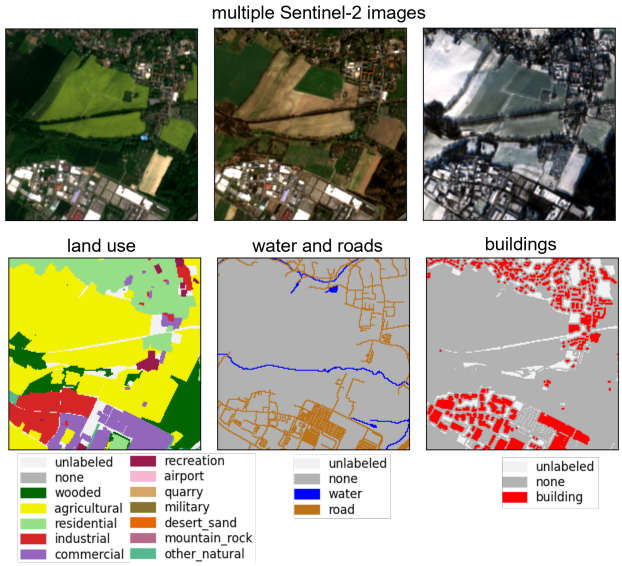

Figure 1. The OpenSentinelMap dataset contains multiple Sentinel-2 3.7 km^2 image crops and corresponding per-pixel labeling derived from OpenStreetMap tags for 137,045 unique locations across the globe.

that support common detection and segmentation tasks. We provide a ready-for-use dataset consisting of satellite images of constant size from the Sentinel-2 platform [3], and per-pixel label images using new land use classes derived from OpenStreetMap (OSM) [9]. OSM tags are collapsed into broad categories useful for land use classification, and polygons are rasterized to match the 10 meter resolution bands of the satellite imagery. Images have been filtered to remove large clouds, while retaining small cloud puffs to support increased model robustness. The dataset as a whole has been biased towards areas of human activity. We envision this data supporting fully supervised, semi-supervised, and unsupervised models.

In this work, we present the proposed Sentinel-2 and OSM dataset, explain in detail the process for acquiring and preprocessing the imagery and OSM labels, train a baseline

neural network for producing embeddings using this data, and provide experimental results showing the usefulness of those embeddings on various down-stream tasks.

2. Related Works

The proposed dataset consists of 1,035,544 publicly available satellite image crops covering 505,202 km^2 distributed across six continents and associated per-pixel class labels for 15 categories derived from OpenStreetMap annotations. Existing public datasets are either smaller in scale, provide annotations narrower in scope, or both.

There exist many remote sensing land-use datasets consisting of high-resolution RGB imagery collected from Google Earth paired with manual annotations crafted by human experts [7, 24]. These datasets are prepared for whole image classification as opposed to dense per-pixel segmentation, and tend to be lacking in either size or global coverage.

Whereas we are providing dense per-pixel label maps to support semantic segmentation, most remote sensing datasets are still focused on single labels to describe an entire image chip. At the coarser resolution of Sentinel-2 imagery this appears insufficient, as image chips are large enough to cover multiple land cover classes. While other datasets offering dense land cover labels do exist [4, 6, 16, 22], they are limited in scope and availability, and rely on high resolution aerial imagery.

Li *et al*. [18] study the influence of seasonal bias in deep learning models trained to detect high-rise buildings from Sentinel-2 imagery across four different seasons. They conclude that fully convolutional networks trained on samples across all seasons can achieve better accuracy than models trained on data from a specific season. Our work makes an effort to evenly distribute images across all seasons.

Due to the size of the proposed dataset, manual per-pixel annotation of all images is not practical; instead, we leverage crowd-sourced OpenStreetMap (OSM) [9] annotations to automatically generate dense labelings for 15 land use classes.

OSM annotations have been used by others to generate land use land cover (LULC) maps [20] and provide ground truth labels for deep neural networks [20, 25]. Other work [15, 19] has mapped OSM tags to classes used by reference datasets such as CORINE Land Cover (CLC) [12], Urban Atlas [13], and GlobeLand30 [1]. The LULC classes used in the proposed dataset were chosen based on a manual grouping of common OSM tags found in the dataset. Many classes can be mapped directly to CLC classes, but some are either more or less specific depending on availability of appropriate OSM tags.

The proposed dataset expands the idea of mapping specific OSM tags to broader land use categories and applies it on a large scale to corresponding Sentinel-2 imagery.

3. Dataset

The proposed dataset consists of publicly available multispectral Sentinel-2 satellite imagery and per-pixel annotations derived from publicly available OpenStreetMap [9] annotations. The source data is processed and filtered into a form convenient for training and testing modern machine learning methods. The Earth is gridded into non-overlapping cells of equal size, and sampled over areas of human activity. Due to the uneven availability of OSM annotations, the labeling in many images is either sparse or nonexistent; unsupervised or semi-supervised methods will therefore be required to make full use of all the available data. Instructions for downloading the dataset are available at https://visionsystemsinc.github.io/open-sentinel-map/.

3.1. Spatial Cells

Each image in the dataset belongs to a single non-overlapping spatial cell of dimension 1920 m $\times$ 1920 m; each Sentinel-2 crop therefore has dimension 192 pixels on a side in the 10m bands, and 96 pixels on a side in the 20m bands. The WGS84 reference ellipsoid is divided into bands of latitude, and those bands are then independently segmented into cells of equal longitudinal width. The bounds of these cells are used when downloading imagery, resulting in multiple image crops covering the same geographic region.

We use VIIRS Nighttime Light (VNL) [11] data to bias our dataset towards areas of human activity. This gives a more general signal than raster data such as the Global Human Settlement Layer [10], which measures settlements specifically by the presence of buildings. The Earth Observation Group (EOG) makes available filtered annual composites of nighttime light data from 2012 to 2020. This data is collected using the Day Night Band (DNB) of the Visible Infrared Imaging Radiometer Suite (VIIRS) on-board satellites within the Joint Polar-orbiting Satellite System. We threshold this data to produce binary human-activity masks for each year. Any pixel with a positive intensity is considered interesting enough for our purposes, and is included in the mask. Each pixel covers 15 arc seconds, which is about 500 meters near the equator.

Spatial cells that do not intersect the binary human-activity mask are removed. Spatial cells completely covered by water according to the Climate Change Initiative Land Cover (CCI-LC) [14] 2020 product are also removed. 137,045 cells are then randomly sampled from those remaining. Their geographic distribution is shown in Figure 2.

The data is split following an 80/10/10 train/val/test format by Military Grid Reference System (MGRS) tile, not by individual spatial cell. This increases the independence of the training, validation, and testing sets by separating them

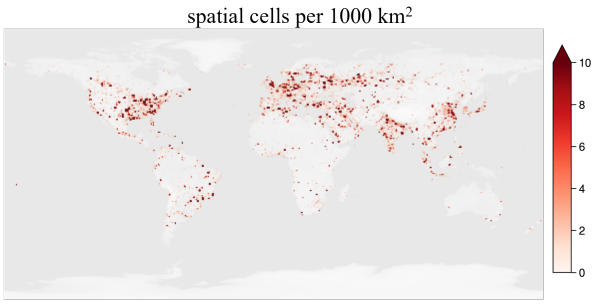

Figure 2. The dataset contains 137,045 spatial cells randomly selected over areas of human activity across the entire Earth.

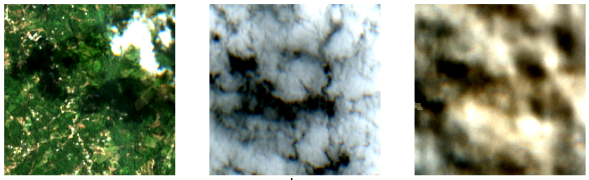

Figure 3. The dataset contains images with minimal cloud cover (a). Images with significant cloud cover are removed from the dataset either by the Sentinel-2 SCL layer (b) or through minimum intensity filtering (c).

at a larger spatial scale.

3.2. Imagery

The European Space Agency (ESA) launched the first satellite of the Copernicus Sentinel-2 satellite in 2015, and the second in 2017 [3]. These two satellites in sun-synchronous orbit are able to completely image the earth every 5 days. The Multi-spectral Instrument (MSI) onboard each satellite captures 13 spectral bands at various resolutions: 10, 20, and 60 meters. The medium resolution, global coverage, and high revisit rate of this data source makes it very promising for tracking land cover and land use changes.

The ESA makes this data freely available, and it is rehosted on the Amazon Web Services (AWS) Registry of Open Data. The images used in our dataset are Level-2A products downloaded from the AWS Sentinel-2 Cloud-Optimized GeoTIFFs (COGs) Simple Storage Service (S3) bucket [2]. Element84 hosts this bucket which replicates the public dataset of Sentinel-2 imagery and converts the file format into COGs, which allows efficient downloading of just a small image chip from the relatively large Sentinel-2 images. The bucket also has a corresponding Spatio-Temporal Asset Catalog (STAC), which was used to search for images covering a specific spatial cell and year, and sort by metadata including total cloud cover.

For each spatial cell, imagery is split over four years spanning the range 2017 - 2020. The earliest Level-2A data available is from 2017. Within each of these combinations of spatial cell and year, a pair of images from two random dates within the year are chosen and downloaded from S3 to local storage. The temporal duration of a year was chosen to allow the full range of seasonal variation, while reducing the probability of true semantic change within the pair as much as possible.

One artifact that may be noticed in a careful inspection of the geographic distribution of spatial cells is that spatial cells tend to cluster into large square tiles. This is due to a detail of our data collection: we process all of the cells within an MGRS tile at once. This is done to minimize STAC searches for efficiency reasons, as the full-sized Sentinel-2 images are fit to MGRS tiles.

To reduce the presence of temporal bias towards times of year with higher probability of cloud-free skies, the images returned by the STAC search are binned by month before being randomly shuffled and processed.

Each Sentinel-2 MSI image downloaded is thus the mapping of a spatial cell into a larger Sentinel-2 image. The downloaded square image chips have a size of 192 pixels at the 10 meter resolution, 96 pixels at 20 meters, and 32 pixels at 60 meters. The intensity values for each band are 32-bit floating point numbers between 0 and 1. These values represent bottom-of-atmosphere surface reflectance.

The Level-2A product includes a Scene Classification Layer (SCL) which contains pre-computed classification maps for ten classes including cloud, cloud shadow, water, and snow. We use this layer to filter image chips as we process the results returned by the STAC search. The SCL classifies individual pixels as cloud or not cloud with different levels of certainty. We err on the side of caution, considering a pixel unusable even when reported to be cloudy with only medium probability. The SCL also attempts to identify cloud shadows, and saturated or defective pixels, which we also consider to be unusable. If enough pixels in an image chip are unusable, we ignore it and move on to the next image returned by the STAC search. Rather than requiring our image chips to be strictly cloud free, we allow up to 25% of the pixels in the image to be cloudy. We acknowledge that no cloud filtering algorithm will ever be completely successful, and we wish to allow networks trained on this data to learn to be robust to the presence of thin and/or sparse clouds in an otherwise useful image.

To further reduce the probability of including heavily clouded images, we add a simple minimum intensity filter to remove brightly saturated scenes caused by false negatives in the scene classification algorithm as demonstrated in Figure 3.

3.3. Annotations

OpenStreetMap (OSM) [9] offers globally available information regarding geospatial entities such as roads, parking lots, and building footprints. Geospatial entities are typically labeled with a variety of tags, for example "building," "parking," "baseball field", and even coarse land use labels. Over 7 million users have contributed to OSM, able to freely create, edit, query and process geodata of various forms.

OSM vector data is rasterized at 10 meter resolution into label images with 15 categories defined by an ontology relating OSM tags to land use classes. The label categories are not necessarily mutually exclusive; for example, the "building", "road", and "water" labels can exist within broader land use regions such as "residential" or "industrial". For this reason, the generated label images are multi-channel; only categories within a channel are assumed to be mutually exclusive. The label images have three channels, as shown in Figure 1: "land use", "water and roads", and "buildings". The individual "land use" categories were chosen manually based on the availability of supporting OSM tags and the ability to visually distinguish the categories in the Sentinel-2 imagery (maximum resolution 10 meters).

The label images are generated using the following process. First, all relevant OSM annotations within the bounds of each spatial cell are downloaded using the OSMNX software package [5]. Next, the ontology mapping individual OSM tags to one or more of the label categories are manually generated. The ontology also contains a precedence value for each category within a channel; in cases where multiple categories map to a single pixel (*e.g.* a bridge over water), the category with the higher precedence value is used. In addition to the positive examples of each category, each channel in the label images contains two additional values: "unlabeled" and "none". The category of pixels with the "unlabeled" value is unknown based on the available OSM tags for that location. A value of "none" indicates that the true category of the pixel is assumed known but does not belong to any of the other categories in the channel. Using the category ontology, a single multi-channel label image for each spatial cell in the dataset is rendered by rasterizing the appropriate OSM annotations. In some cases, individual tags are dilated after rasterization; for example, any pixel within 30 meters of the "building: house" OSM tag is labeled as "residential". For additional details concerning the mapping of OSM tags to categories, please see the `osm_categories.json` file distributed with the dataset.

3.3.1 Label Statistics

While OSM data is available globally, it is not of uniform quality and completeness. Statistics regarding the density of annotations in general and of individual categories are shown in Figure 4. The density of labels across each spatial cell follows a bi-modal distribution with peaks at either extreme of the scale; 11% of spatial cells contain less than 1% of pixels labeled, while 12% of spatial cells contain over 90% of pixels labeled with at least one of the 15 categories.

The geospatial distribution of sparsely and densely labeled spatial cells is also non-uniform; Densely labeled spatial cells are most common in Western Europe and in large cities, as shown in Figure 5. Sparsely labeled (or empty) spatial cells are common everywhere except for Western Europe.

Finally, the prevalence of the individual categories is also distributed non-uniformly. Approximately 88% of spatial cells contain at least one labeled road, while the "quarry" label is present in just under 3% of spatial cells. After "road", the "water" and "building" labels are the next most common labels (found in 57% and 37% of images, respectively). When measured on a per-pixel (as opposed to per-image) basis, the most common labels are those typically associated with large tracts of land such as "wooded" and "agricultural" (approximately 8.7% and 7.3% of all pixels in the dataset, respectively.)

3.3.2 Label Inaccuracies

The scale and diversity of the OSM-derived label images provide significant value to the remote sensing and machine learning community, including the ability to train high-quality landuse, building, and road detection in globally available imagery using simple methods as demonstrated in Section 5. However, the nature of the source data leads to potential inaccuracies that practitioners using the dataset should be aware of. In addition to the uneven distribution of labels (Section 3.3.1), georegistration, rasterization, and temporal errors also exist in the data. Georegistration and rasterization errors are caused by annotations that are incorrectly placed or made as linear paths ("ways") as opposed to polygons. In the latter case, the annotations are rasterized by "turning on" each pixel that the path passes through optionally followed by a dilation, depending on the associated tag. For example, based on the assumed width of the road, the OSM tag "highway: trunk" is dilated after rasterization, but "highway: residential" is not. Rivers of significant width are typically annotated with polygons providing the full boundary, but are occasionally annotated using linear paths only. In these cases, no assumption about the width of the river is made; the annotation is left as a one-pixel wide path in the label image. In addition to geospatial inaccuracies, OSM data may also be misaligned temporally with the included Sentinel-2 images, particularly in regions with large amounts of active construction. The OSM data was downloaded in early 2022, while the Sentinel-2 images span the range 2017 through 2020. While it is possible to

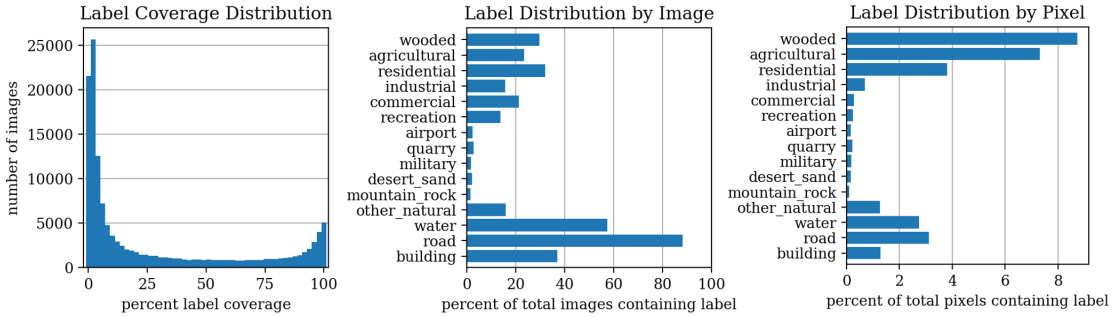

Figure 4. Per-image label coverage exhibits a bi-modal distribution (left); many images contain little or no labels, and many images are nearly fully labeled. Prevalence of individual labels varies significantly in both image count (middle) and total pixel count (right).

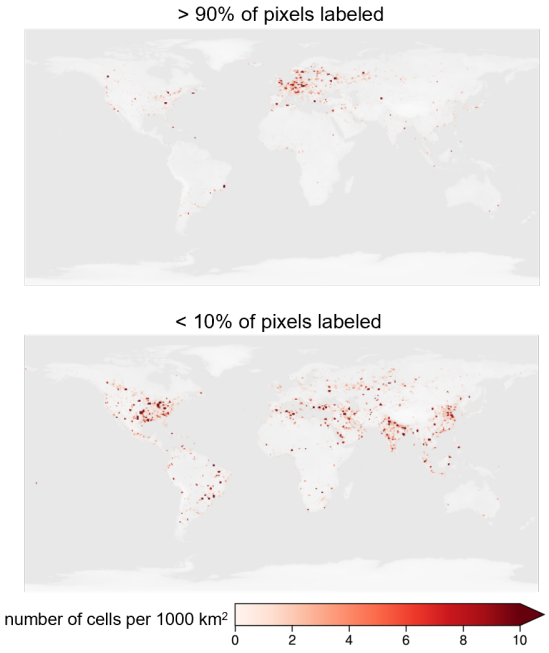

Figure 5. OSM-derived labels are available globally but are not distributed uniformly. Images with dense label coverage are common mainly in Western Europe and in large cities.

retrieve historical snapshots of OSM data, it is generally not clear if new annotations are added to fill in missing data over a static region or to indicate real changes. For this reason, a single label image is generated for each spatial cell using the most recently available OSM annotations, with no attempt to model historical changes on the ground.

4. Embedding Network Architecture

To demonstrate the utility of OpenSentinelMap and establish baseline performance, we present an encoder network that learns rich, semantic embeddings for Sentinel-2 imagery by using per-pixel OSM classification as a proxy learning task.

The encoder backbone is a modified 3-stage HRNet [21] that produces dense, per-pixel embeddings which can be utilized in downstream tasks (Fig. 6). HRNet was chosen for its performance and sharpness on high-resolution semantic segmentation tasks, because pixel-wise location accuracy is important for remote sensing imagery. The desire for faster inference on large-scale satellite imagery drove the decision to generate dense embeddings for every pixel in a fully convolutional manner. The network is trained using temporally differing image pairs of the same Sentinel-2 spatial cell (Section 3.1) (*i.e.* the same location at two different dates). The pairwise training strategy is adopted to encourage imagery that may contain seasonal and weather variance to produce similar embeddings.

4.1. Training Data

The training data consists of the four Sentinel-2 10m bands (near infrared, red, green, and blue bands), the six 20m bands (vegetation red edge bands, and SWIR bands), and the corresponding OSM class labels from OpenSentinelMap. Each spatial cell (Section 3.1) is 192 pixels on each side for the 10m bands (1920 m.) and 96 pixels on each side for the 20m band (1920 m.), for resulting input sizes of 192x192x4 and 96x96x6 respectively. We elect to exclude the 60m Sentinel-2 bands from training as they are mainly utilized in the pre-processing procedure for producing the Level-2A imagery and are too coarse for several categories such as buildings. The dataset is split into approximately 110k spatial cells for training, 15k spatial cells for validation, and 11k spatial cells for testing (for experimental results on the testing set, see Section 5.1).

4.2. Modification for Sentinel-2 Imagery

Instead of only utilizing the 10m bands or upsampling the 20m bands to match the input size of the 10m bands, we choose to learn individual features at each of the original input resolutions and feed those feature maps into their respective scales in the HRNet architecture. Therefore, the

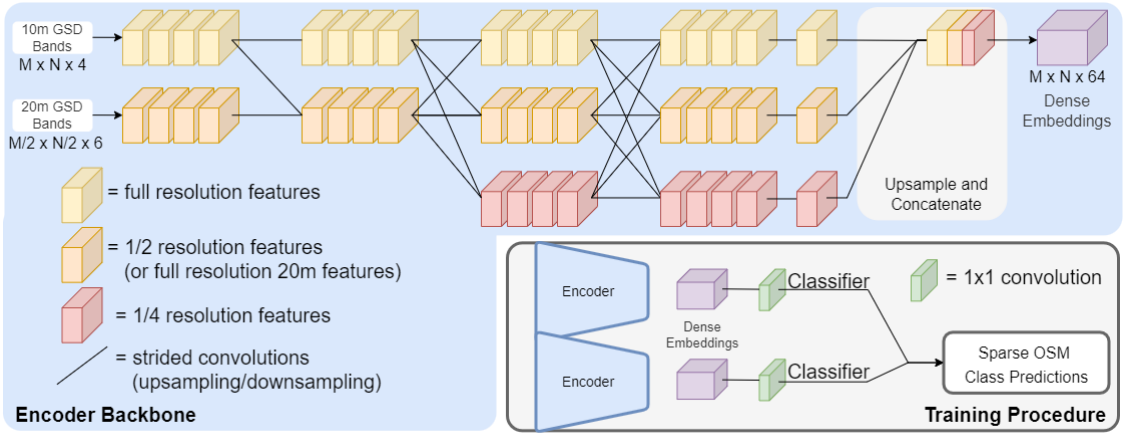

Figure 6. Illustrations of the encoder network architecture and the pair-wise training procedure. Yellow, orange, and red blocks indicate HRNet convolutional blocks and features at specific resolutions and purple blocks represent the produced dense embeddings. In the training procedure, each blue encoder block represents an entire encoder backbone and green blocks are the same 1x1 convolutional classifier.

10m bands are input into the typical HRNet input convolution block, and the network is modified by appending an additional input convolution block for the 20m bands that feeds forward into the $1/2$ resolution feature branch of the network (Fig. 6).

The standard HRNet network has a larger receptive field at maximum resolution than the image size of the 10m bands for an entire spatial cell. To prevent the network from seeing heavily padded images during training, the receptive field was reduced by removing a stage of pooling and strided convolution operations at the smallest feature resolution. Batchnorm layers are replaced with Group Normalization [23] to facilitate training with small batch size.

4.3. Network Training

The encoder network and OSM classification head were trained end-to-end for 225 epochs using an SGD optimizer ($LR = 0.01$; $momentum = 0.9$) and a linear warm-up cosine annealing learning rate scheduler (10 warm-up epochs; $LR_{\text{start}} = 0.00$; $LR_{\text{end}} = 0.0001$). Training imagery was standardized with mean-variance normalization and augmented using random rotations and mirroring. Due to the sparse nature of the OSM ground truth class labels, we only calculate and average a binary cross-entropy loss where either a true positive or true negative ("none" labeled) pixel exists in the OSM label image (i.e. "unlabeled" pixels do not contribute to the loss).

5. Experiments

We quantitatively analyse the performance of the trained network described above in Section 4 by evaluating on the testing set of OpenSentinelMap as well as leveraging embeddings learned using OpenSentinelMap to perform downstream classification on two existing datasets: EuroSAT [17] and Functional Map of the World [8]. The goal of the EuroSAT experiment is to determine if embeddings learned from the OpenSentinelMap labels can be used to train a coarse land cover classifier using a small training set. The goal of the FMoW experiment is to determine if our embeddings contain enough detailed semantic information to support fine-grained classifications.

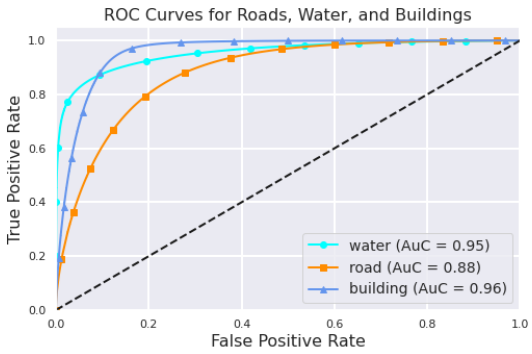

Figure 7. ROC Curves and Area under those curves (AuC) for the road, water, and building classes in the OSM test dataset.

5.1. Test Set Results

For evaluation on the test set of OpenSentinelMap, we simply use the encoder network and the pre-trained OSM per-pixel classification head. Semantic segmentation maps are predicted for each of the spatial cells in the test set and compared to the ground truth OSM label images.

Figure 7 shows ROC curves as well as area under those curves (AuC) for the road, water, and building classes which

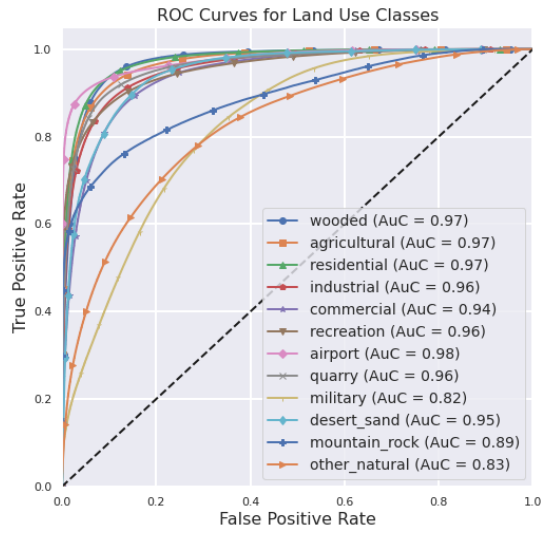

Figure 8. ROC Curves and Area under those curves (AuC) for the 12 land use classes in the OSM test dataset.

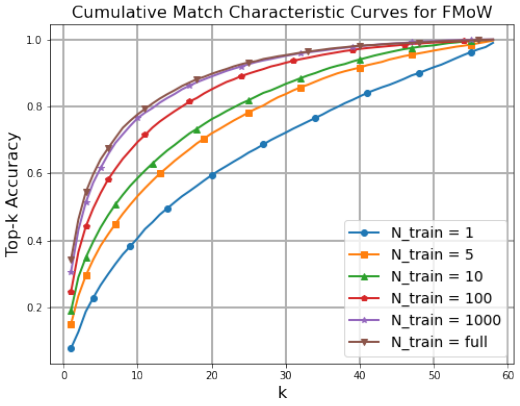

Figure 9. CMC curves showing the top-k accuracy for the 10m FMoW classifier for multiple values of N_{train}.

are not mutually exclusive with the 12 land use classes. Figure 8 shows ROC curves and AuCs for the remaining 12 land use classes. Figure 10 shows network predictions for an example image along with OSM ground truth label images.

5.2. EuroSAT

The EuroSAT [17] dataset consists of 64x64 pixel patches of Level-1A (L1A) Sentinel-2 MSI imagery and includes 10 different land use classes. In total, there are 27k labeled and geo-referenced L1A images across all classes. Because the network is trained with Level-2A Sentinel-2 imagery (L2A), the geographic boundaries for each EuroSAT image were used to resample imagery from L2A data. This L2A EuroSAT dataset will be published alongside OpenSentinelMap.

In order to evaluate the number of training samples needed for training, the EuroSAT data is split into training sets of multiple sizes. These splits are either fixed sizes (where N_{train} is the number of samples per class) or percentage-based splits of training/testing data (i.e. $80/20$ is 80% training and 20% testing). For fixed training splits, a fixed testing split of 1000 samples per class is used.

First, the training data is fed into the encoder (Sec. 4) to generate a set of embeddings for each image. The resulting embeddings are cropped down to the original EuroSAT image extents (64x64) and detached from the encoder network (i.e. the backward pass will not update the encoder). A small 3-layer Convolutional Neural Network (CNN) classification head is then trained on top of the cropped and detached embeddings using the EuroSAT land use labels.

Results for each of the values of N_{train} (as well as percentage-based splits to compare to [17]) are shown in Table 1. Our semantically rich embeddings allow a small classification head with a small number of the two baselines' trainable parameters (approximately 1% of ResNet-50 and 3% of GoogleNet) to achieve strong results with a small number of labelled images. Above 600 samples per classes (30/70 split) we start to see diminishing returns in performance compared to the baselines.

5.3. FMoW

The Functional Map of the World (FMoW) [8] dataset consists of annotations for a variety of unique building and land use classes. In total, there are 130k annotations with at least one labeled bounding box and one of 62 categories.

Sentinel-2 imagery is of a much lower resolution (10m per pixel) than the original imagery included in the FMoW dataset (≈0.5m per pixel). Because the dataset is comprised of classes that have varying spatial extents, classifying relatively small objects in the FMoW dataset is a challenging evaluation task for the network.

We construct a dataset and use it to train a simple classifier in a similar way to the EuroSAT protocol described in Section 5.2. Initially, the FMoW ground truth annotations are used to resample L2A imagery. Our resampling ignores the *construction site*, *flooded road*, and *impoverished settlement* classes leaving a total of 59 categories in the dataset. We postulate that these named categories are not suitable for single-view classification due to their temporal nature. The resampled Sentinel-2 FMoW dataset will also be published alongside OpenSentinelMap.

Multiple training splits of size N_{train} samples per class are selected along with the FMoW validation and testing sets. The resampled FMoW images are fed through the encoder network to produce embeddings. Because [8] use 244x244 pixel imagery in their classification procedure (approximately 122m per side), the embeddings are center

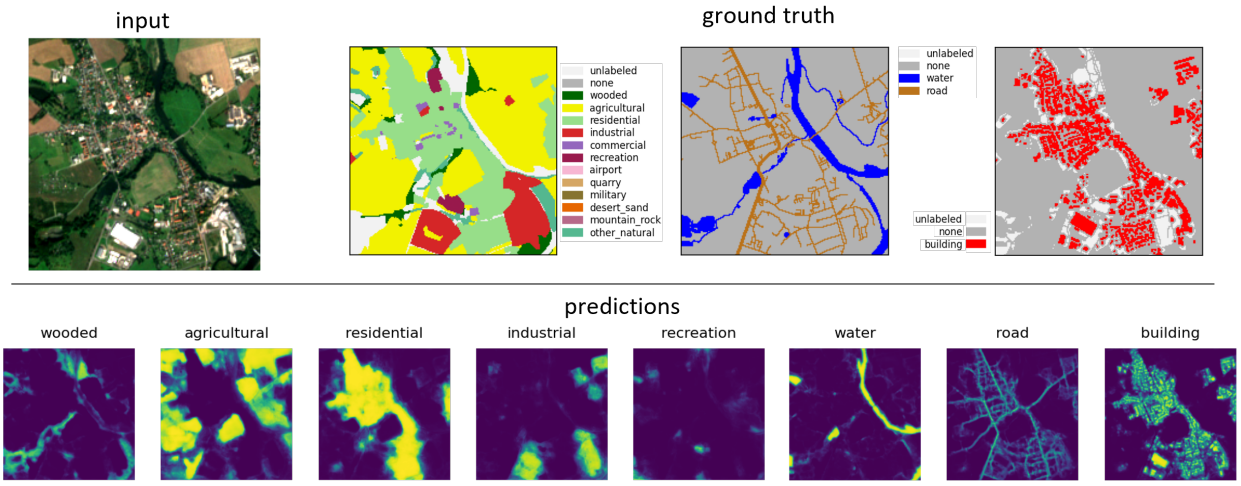

Figure 10. OSM-based ground truth images (top) are used to evaluation per-pixel probabilities for each of the categories predicted by the trained network (bottom). Note that only the 8 prediction classes out of 15 with non-negligible probabilities are shown here.

Method	1	5	10	100	10/90	20/80	30/70	40/60	50/50	60/40	70/30	80/20	90/10
ResNet-50 [17]	×	×	×	×	75.06	88.53	93.75	94.01	94.45	95.26	95.32	96.43	96.37
GoogleNet [17]	×	×	×	×	77.37	90.97	90.57	91.62	94.96	95.54	95.70	96.02	96.17
Ours	57.56	74.58	81.66	91.48	92.45	93.10	93.62	93.80	93.95	93.71	93.80	94.90	94.70

Table 1. Top-1 Accuracy for the EuroSAT classifier trained on top of the learned embeddings. Accuracy is reported for single-shot, few-shot, and percentage-based train / test splits comparable with baseline methods.

cropped to 16x16 pixels (160m per side) to approximate the same footprint used in the classification experiments. Finally, these cropped embeddings are detached and used to train a 3-layer CNN classification head on the 59 FMoW categories.

The Cumulative Match Characteristic (CMC) curve for multiple values of N_{train} is shown in Figure 9. While performance is significantly below methods that use high resolution imagery, we demonstrate that reasonable baseline performance is possible using 10 meter resolution imagery.

6. Conclusion

We have presented a new dataset consisting of globally sampled multi-spectral Sentinel-2 imagery and corresponding per-pixel labels for 15 non-mutually exclusive semantic categories based on OpenStreetMap annotations.

The samples were selected without regard to availability or quality of annotation data, making the dataset an ideal candidate for unsupervised and semi-supervised approaches. We leave the development and testing of these approaches for future work, and instead demonstrate that the quality and quantity of annotations is sufficient for training a high-quality baseline method using a standard fully-supervised approach.

In addition, we demonstrate that the internal representation learned by the supervised method is sufficient for training downstream classifiers of fine-grained categories not explicitly labeled in the training set. Despite its significant size, the proposed dataset is built using only a small fraction of the total available Sentinel-2 and OpenStreetMap data; future work will investigate the question of how much data is required before reaching the point of diminishing returns on downstream segmentation, detection, and classification tasks.

Acknowledgement

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via 2021-2011000004. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

[1] Global land cover. http://www.globallandcover.com/. 2

[2] Element 84. Sentinel-2 cloud-optimized geotiffs. https://registry.opendata.aws/sentinel-2-l2a-cogs/, 2022. 3

[3] European Space Agency. Sentinel-2 mission guide. https://sentinel.esa.int/web/sentinel/missions/sentinel-2, 2022. 1, 3

[4] Seyed Majid Azimi, Corentin Henry, Lars Sommer, Arne Schumann, and Eleonora Vig. Skyscapes fine-grained semantic understanding of aerial scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7393–7403, 2019. 2

[5] Geoff Boeing. Osmnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Computers, Environment and Urban Systems*, 65:126–139, Sep 2017. 4

[6] Adrian Boguszewski, Dominik Batorski, Natalia Ziemba-Jankowska, Tomasz Dziedzic, and Anna Zambrzycka. Landcover.ai: Dataset for automatic mapping of buildings, woodlands, water and roads from aerial imagery. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1102–1110, 2021. 2

[7] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *CoRR*, abs/1703.00121, 2017. 2

[8] Gordon A. Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. *CoRR*, abs/1711.07846, 2017. 6, 7

[9] OpenStreetMap contributors. Openstreetmap. https://www.openstreetmap.org, 2022. 1, 2, 4

[10] Christina Corbane, Vasileios Syrris, Filip Sabo, Panagiotis Politis, Michele Melchiorri, Martino Pesaresi, Pierre Soille, and Thomas Kemper. Convolutional neural networks for global human settlements mapping from sentinel-2 satellite imagery. *Neural Computing and Applications*, 33(12):6697–6720, 2021. 2

[11] Christopher D Elvidge, Mikhail Zhizhin, Tilottama Ghosh, Feng-Chi Hsu, and Jay Taneja. Annual time series of global viirs nighttime lights derived from monthly averages: 2012 to 2019. *Remote Sensing*, 13(5):922, 2021. 2

[12] ESA. Corine land cover. https://land.copernicus.eu/pan-european/corine-land-cover. 2

[13] ESA. Urban atlas. https://land.copernicus.eu/local/urban-atlas. 2

[14] ESA. Land cover cci product user guide version 2. tech. rep. maps.elie.ucl.ac.be/CCI/viewer/download/ESACCI-LC-Ph2-PUGv2_2.0.pdf, 2017. 2

[15] Cidália C Fonte, Joaquim Patriarca, Ismael Jesus, and Diogo Duarte. Automatic extraction and filtering of openstreetmap data to generate training datasets for land use land cover classification. *Remote Sensing*, 12(20):3428, 2020. 2

[16] Markus Gerke, Franz Rottensteiner, Jan Wegner, and Gunho Sohn. Isprs semantic labeling contest. 09 2014. 2

[17] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2019. 6, 7, 8

[18] Liwei Li, Jinming Zhu, Gang Cheng, and Bing Zhang. Detecting high-rise buildings from sentinel-2 data based on deep learning method. *Remote Sensing*, 13(20), 2021. 2

[19] J Patriarca, CC Fonte, J Estima, J-P de Almeida, and A Cardoso. Automatic conversion of osm data into lulc maps: comparing foss4g based approaches towards an enhanced performance. *Open Geospatial Data, Software and Standards*, 4(1):1–19, 2019. 2

[20] Michael Schultz, Janek Voss, Michael Auer, Sarah Carter, and Alexander Zipf. Open land cover from openstreetmap and remote sensing. *International Journal of Applied Earth Observation and Geoinformation*, 63:206–213, 2017. 2

[21] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *TPAMI*, 2019. 5

[22] Shenlong Wang, Min Bai, Gellert Mattyus, Hang Chu, Wenjie Luo, Bin Yang, Justin Liang, Joel Cheverie, Sanja Fidler, and Raquel Urtasun. Torontocity: Seeing the world with a million eyes. *arXiv preprint arXiv:1612.00423*, 2016. 2

[23] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 6

[24] Gui-Song Xia, Jingwen Hu, Fan Hu, Baoguang Shi, Xiang Bai, Yanfei Zhong, and Liangpei Zhang. AID: A benchmark dataset for performance evaluation of aerial scene classification. *CoRR*, abs/1608.05167, 2016. 2

[25] Wenzhi Zhao, Yanchen Bo, Jiage Chen, Dirk Tiede, Thomas Blaschke, and William J. Emery. Exploring semantic elements for urban scene recognition: Deep integration of high-resolution imagery and openstreetmap (osm). *ISPRS Journal of Photogrammetry and Remote Sensing*, 151:237–250, 2019. 2