

Functional Map of the World

Gordon Christie¹

Neil Fendley¹

James Wilson²

Ryan Mukherjee¹

¹The Johns Hopkins University Applied Physics Laboratory

²DigitalGlobe

{[gordon.christie](mailto:gordon.christie@jhuapl.edu), [neil.fendley](mailto:neil.fendley@jhuapl.edu), [ryan.mukherjee](mailto:ryan.mukherjee@jhuapl.edu)}@jhuapl.edu

james.wilson@digitalglobe.com

Abstract

We present a new dataset, *Functional Map of the World* (fMoW), which aims to inspire the development of machine learning models capable of predicting the functional purpose of buildings and land use from temporal sequences of satellite images and a rich set of metadata features. The metadata provided with each image enables reasoning about location, time, sun angles, physical sizes, and other features when making predictions about objects in the image. Our dataset consists of over 1 million images from over 200 countries¹. For each image, we provide at least one bounding box annotation containing one of 63 categories, including a “false detection” category. We present an analysis of the dataset along with baseline approaches that reason about metadata and temporal views. Our data, code, and pretrained models have been made publicly available.

1. Introduction

Satellite imagery presents interesting opportunities for the development of object classification methods. Most computer vision (CV) datasets for this task focus on images or videos that capture brief moments [24, 20]. With satellite imagery, temporal views of objects are available over long periods of time. In addition, metadata is available to enable reasoning beyond visual information. For example, by combining temporal image sequences with timestamps, models may learn to differentiate office buildings from multi-unit residential buildings by observing whether or not their parking lots are full during business hours. Models may also be able to combine certain metadata parameters with observations of shadows to estimate object heights. In addition to these possibilities, robust models must be able to generalize to unseen areas around the world that may include different building materials and unique architectural styles.

Enabling the aforementioned types of reasoning requires a large dataset of annotated and geographically diverse

¹fMoW contains 1,047,691 images covering 207 of the total 247 ISO Alpha-3 country codes.

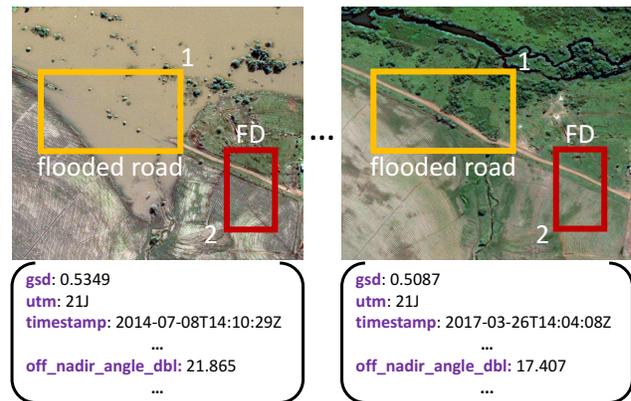


Figure 1: In fMoW, temporal sequences of images, multispectral imagery, metadata, and bounding boxes are provided. In this example, if we only look inside the yellow box in the right image, we will only see a road and vegetation. On the other hand, if we only see the water in the left image, then we will potentially predict this to be a lake. However, by observing both views of this area, we can now reason that this sequence contains a flooded road.

satellite images. In this work, we present our efforts to collect such a dataset, entitled *Functional Map of the World* (fMoW). fMoW has several notable features, including global diversity, a variable number of temporal images per scene, multispectral imagery, and metadata associated with each image. The task posed for our dataset falls in between object detection and classification. That is, for each temporal sequence of images, at least one bounding box is provided that maps to one of 63 categories, including a “false detection” (FD) category that represents content not characterized by the other 62 categories. These boxes are intended to be used as input to a classification algorithm. Figure 1 shows an example.

Collecting a dataset such as fMoW presents some interesting challenges. For example, one consideration would be to directly use crowdsourced annotations provided by OpenStreetMap² (OSM). However, issues doing so include

²<https://www.openstreetmap.org>

inconsistent, incorrect, and missing annotations for a large percentage of buildings and land use across the world. Moreover, OSM may only provide a single label for the current contents of an area, making it difficult to correctly annotate temporal views. Another possibility is to use the crowd to create annotations from scratch. However, annotating instances of a category with no prior information is extremely difficult in a large globally-diverse satellite image dataset. This is due, in part, to the unique perspective that satellite imagery offers when compared with ground-based datasets, such as ImageNet [24]. Humans are seldom exposed to aerial viewpoints in their daily lives. As such, objects found in satellite images tend to be visually unfamiliar and difficult to identify. Buildings can also be repurposed throughout their lifetime, making visual identification even more difficult. For these reasons, we use a multi-phase process that combines map data and crowdsourcing.

Another problem for fMoW is that annotating every instance of a category is made very difficult by the increased object density of certain categories. For example, single-unit residential buildings often occur in dense clusters alongside other categories, where accurately discriminating and labeling every building would be very time-consuming. To address this shortcoming, we propose providing bounding boxes as algorithm input, unlike a typical detection dataset and challenge where bounding boxes are expected as output. This circumvents full image annotation issues that stem from incomplete map data and visual unfamiliarity. As a result, data collection could focus on global diversity and annotations could be limited to a small number of high-confidence category instances per image.

Our contributions are summarized as follows: (1) We provide the largest publicly available satellite dataset containing bounding box annotations, multispectral imagery, metadata, and revisits. This enables joint reasoning about images and metadata, as well as long-term temporal reasoning for areas of interest. (2) We present methods based on CNNs that exploit the novel aspects of our dataset, with performance evaluation and comparisons, which can be applied to similar problems in other application domains. Our code, data, and pretrained models have all been publicly released³. In the following sections, we provide an analysis of fMoW and baseline methods for the task.

As an aside, in addition to collecting and publishing fMoW, a public prize challenge⁴ was organized around the dataset. It ran from Sep. 14 - Dec. 31 2017. The top 3 participants have open-sourced their solutions on the fMoW GitHub page. These methods, as well as the baseline, were developed using the publicly available data. However, all data, including the sequestered data used for final testing, is now publicly available.

³<https://github.com/fMoW>

⁴<https://www.iarpa.gov/challenges/fmow.html>

2. Related Work

While large datasets are nothing new to the vision community, they have typically focused on first-person or ground-level imagery [24, 20, 2, 10, 11, 9, 19]. This is likely due in part to the ease with which this imagery can be collected and annotated. Recently, there have been several, mostly successful, attempts to leverage techniques that were founded on first-person imagery and apply them to remote sensing data [15, 21, 30]. However, these efforts highlight the research gap that has developed due to the lack of a large dataset to appropriately characterize the problems found in remote sensing. We now highlight several of these areas where we believe fMoW can make an impact.

Reasoning Beyond Visual Information Many works have extended CV research to simultaneously reason about other modules of perception [3, 16, 23, 12, 4]. In this work, we are interested in supporting joint reasoning about temporal sequences of images and associated metadata features. One of these features is UTM zone, which provides location context. In a similar manner, [26] shows improved image classification results by jointly reasoning about GPS coordinates and images, where several features are extracted from the coordinates, including high-level statistics about the population. Although we use coarser location features (UTM zones) than GPS in this work, we do note that using similar features would be an interesting study. GPS data for fMoW imagery was also made publicly available after the end of the prize challenge.

Multi-view Classification Satellite imagery offers a unique and somewhat alien perspective on the world. Most structures are designed for recognition from ground level. As such, it can be difficult, if not impossible, to identify functional purpose from a single overhead image. One of the ways fMoW attempts to address this issue is by providing multiple temporal views of each object, when available. Along these lines, several works in the area of video processing have been able to build upon advancements in single image classification [17, 8, 32] to create networks capable of extracting spatio-temporal features. These works may be a good starting point, but it is important to keep in mind the vastly different temporal resolution on which these datasets operate. For example, the YouTube-8M dataset [2] contains videos with 30 frames per second. For satellites, it is not uncommon for multiple days to pass before they can image the same location, and possibly months before they can get an unobstructed view.

Perhaps the most similar work to ours in terms of temporal classification is PlaNet [28]. They pose the image localization task as a classification problem, where photo albums are classified as belonging to a particular bucket that bounds an area on the globe. We use a similar approach in one of our baseline methods.

Remote Sensing Datasets One of the earliest annotated

satellite datasets similar to fMoW is the UC Merced Land Use Dataset, which offers 21 categories and 100 images per category with roughly 30cm resolution and image sizes of 256x256 [31]. Another recent dataset similar to fMoW is TorontoCity [27], which includes aerial imagery captured during different seasons in the greater Toronto area. While they present several tasks, the two that are similar to land-use classification are zoning classification and segmentation (e.g., residential, commercial). Datasets have also been created for challenges centered around semantic segmentation, such as the IEEE GRSS Data Fusion Contests [6] and the ISPRS 2D Semantic Labeling Contest [1].

SpaceNet [7], a recent dataset that has received substantial attention, contains both 30cm and 50cm data of 5 cities. While it mainly includes building footprints, point of interest (POI) data was recently released into SpaceNet that includes locations of several categories within Rio de Janeiro. Other efforts have also been made to label data from Google Earth, such as the AID [29] (10,000 images, 30 categories) and NWPU-RESISC45 (31,500 images of 45 categories) [5] datasets. In comparison, fMoW offers 1,047,691 images of 63 categories, and includes associated metadata, temporal views, and multispectral data, which are not available from Google Earth.

3. Dataset Collection

Prior to the dataset collection process for fMoW, a set of categories had to be identified. Based on our target of 1 million images, collection resources, plan to collect temporal views, and discussions with researchers in the CV community, we set a goal of including between 50 and 100 categories. We searched sources such as the OSM Map Features⁵ list and NATO Geospatial Feature Concept Dictionary⁶ for categories that highlight some of the challenges discussed in Section 2. For example, “construction site” and “impoverished settlement” are categories from fMoW that may require temporal reasoning to identify, which presents a unique challenge due to temporal satellite image sequences typically being scattered across large time periods. We also focused on grouping categories according to their functional purpose to encourage the development of approaches that can generalize. For example, by grouping recreational facilities (e.g., tennis court, soccer field), algorithms would hopefully learn features common to these types of facilities and be able to recognize other recreational facilities beyond those included in the dataset (e.g., rugby fields). This also helps avoid issues related to label noise in the map data.

Beyond research-based rationales for picking certain categories, we had some practical ones as well. Before categories could be annotated within images, we needed to find

locations where we have high confidence of their existence. This is where maps play a crucial role. “Flooded road”, “debris or rubble”, and “construction site” were the most difficult categories to collect since open source data does not generally contain temporal information. However, with more careful search procedures, reuse of data from humanitarian response campaigns, and calculated extension of keywords to identify categories even when not directly labeled, we were able to collect temporal stacks of imagery that contained valid examples.

All imagery used in fMoW was collected from the DigitalGlobe constellation⁷. Images were gathered in pairs, consisting of 4-band or 8-band multispectral imagery in the visible to near-infrared region, as well as a pan-sharpened RGB image that represents a fusion of the high-resolution panchromatic image and the RGB bands from the lower-resolution multispectral image. 4-band imagery was obtained from either the QuickBird-2 or GeoEye-1 satellite systems, whereas 8-band imagery was obtained from WorldView-2 or WorldView-3.

More broadly, fMoW was created using a three-phase workflow consisting of location selection, image selection, and bounding box creation. The location selection phase was used to identify potential locations that map to our categories while also ensuring geographic diversity. Potential locations were drawn from several Volunteered Geographic Information (VGI) datasets, which were conflated and curated to remove duplicates. To ensure diversity, we removed neighboring locations within a specified distance (typically 500m) and set location frequency caps for categories that have severely skewed geographic distributions. These two factors helped reduce spatial density while also encouraging the selection of locations from disparate geographic areas. The remaining locations were then processed using DigitalGlobe’s GeoHIVE⁸ crowdsourcing platform. Members of the GeoHIVE crowd were asked to validate the presence of categories in satellite images, as shown in Figure 2.

The image selection phase comprised of a three-step process, which included searching the DigitalGlobe satellite imagery archive, creating image chips, and filtering out cloudy images. Approximately 30% of the candidate images were removed for being too cloudy. DigitalGlobe’s IPE Data Architecture Highly-available Object-store service was used to process imagery into pan-sharpened RGB and multispectral image chips in a scalable fashion. These chips were then passed through a CNN architecture to classify and remove any undesirable cloud-covered images.

Finally, images that passed through the previous two phases were sent to a curated and trusted crowd for bounding box annotation. This process involved a separate in-

⁵https://wiki.openstreetmap.org/wiki/Map_Features

⁶https://portal.dgiwg.org/files/?artifact_id=8629

⁷<https://www.digitalglobe.com/resources/satellite-information>

⁸<https://geohive.digitalglobe.com>



Figure 2: Sample image of what a GeoHIVE user might see while validating potential fMoW dataset features. Instructions can be seen in the top-left corner that inform users to press the ‘1’, ‘2’, or ‘3’ keys to validate existence, non-existence, or cloud obscuration of a particular object.

interface from the first phase, where crowd users were asked to draw bounding boxes around the category of interest in each image and were provided some category-specific guidance for doing so. The resulting bounding boxes were then graded by a second trusted crowd to assess quality. The trusted crowd includes individuals from universities and elsewhere that have a strong relationship with DigitalGlobe or the labeling campaigns they have conducted. In total, 642 unique GeoHIVE users required a combined total of approximately 2,800 hours to annotate category instances for fMoW.

Even after multiple crowd validation procedures and implementing programmatic methods for ensuring geographic diversity, there were several categories that contained some bias. For example, the “wind farm” category does not contain very many examples from the United States, even though the initial location selection phase returned 1,938 viable locations from the United States. Many of these “wind farm” instances were invalidated by the crowd, likely due to the difficulty of identifying tall, thin structures in satellite imagery, particularly when the satellite image is looking straight down on the tower. The “barn”, “construction site”, “flooded road”, and “debris or rubble” categories are also examples that contain some geographic bias. In the case of the “barn” category, the bias comes from the distribution of “barn” tags in OSM, which are predominately located in Europe, whereas the other three categories contain geographic bias as a result of the more complex feature selection process, mentioned earlier, that was required for these categories. FD boxes were included to mitigate this bias. When they are present in an image, algorithms are forced to use the imagery to accurately make predictions, as there may be two boxes with different labels that share similar

metadata features.

The following provides a summary of the metadata features included in our dataset, as well as any preprocessing operations that are applied before input into the baseline methods:

- **UTM Zone** One of 60 UTM zones and one of 20 latitude bands are combined for this feature. We convert these values to 2 coordinate values, each between 0 and 1. This is done by taking the indices of the values within the list of possible values and then normalizing. While GPS data is now publicly available, it was withheld during the prize challenge to prevent participants from precisely referencing map data.
- **Timestamp** The year, month, day, hour, minute, second, and day of the week are extracted from the timestamp and added as separate features. The timestamp provided in the metadata files is in Coordinated Universal Time (UTC).
- **GSD** Ground sample distance, measured in meters, is provided for both the panchromatic and multispectral bands in the image strip. The panchromatic images used to generate the pan-sharpened RGB images have higher resolution than the MSI, and therefore have smaller GSD values. These GSD values, which describe the physical sizes of pixels in the image, are used directly without any preprocessing.
- **Angles** These identify the angle at which the sensor is imaging the ground, as well as the angular location of the sun with respect to the ground and image. The following angles are provided:
 - *Off-nadir Angle* Angle in degrees (0-90°) between the point on the ground directly below the sensor and the center of the image swath.
 - *Target Azimuth* Angle in degrees (0-360°) of clockwise rotation off north to the image swath’s major axis.
 - *Sun Azimuth* Angle in degrees (0-360°) of clockwise rotation off north to the sun.
 - *Sun Elevation* Angle in degrees (0-90°) of elevation, measured from the horizontal, to the sun.
- **Image+box sizes** The pixel dimensions of the bounding boxes and image size, as well as the fraction of the image width and height that the boxes occupy, are added as features.

A full list of metadata features and their descriptions can be found in the supplement.

4. Dataset Analysis

Here we provide some statistics and analysis of fMoW. Two versions of the dataset are publicly available:

- **fMoW-full** The full version of the dataset includes pan-sharpened RGB images and 4/8-band multispectral images (MSI), which are both stored in TIFF for-

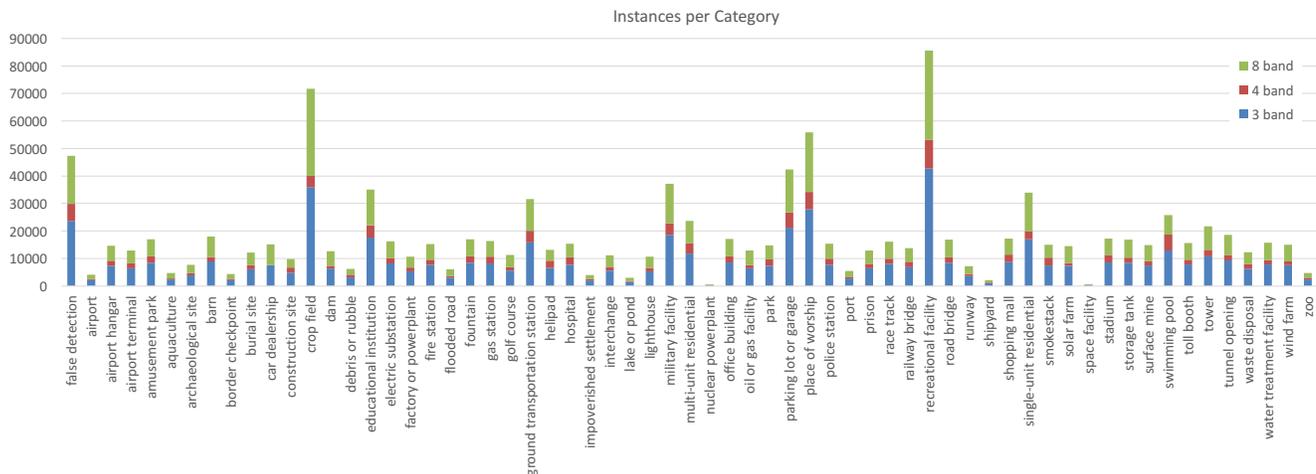


Figure 3: This shows the total number of instances for each category (including FD) in fMoW across different number of bands. These numbers include the temporal views of the same areas. $fMoW-full$ consists of 3 band imagery (pan-sharpened RGB), as well as 4 and 8 band imagery. In $fMoW-rgb$, the RGB channels of the 4 and 8 band imagery are extracted and saved as JPEG images.

mat. Pan-sharpened images are created by “sharpening” lower-resolution MSI using higher-resolution panchromatic imagery [22]. All pan-sharpened images in $fMoW-full$ have corresponding MSI, where the metadata files for these images are nearly identical.

- $fMoW-rgb$ An alternative JPEG compressed version of the dataset, which is provided due to the large size of $fMoW-full$. For each pan-sharpened RGB image we simply perform a conversion to JPEG. For MSI images, we extract the RGB channels and save them as JPEGs.

For all experiments presented in this paper, we use $fMoW-rgb$. We also exclude RGB-extracted versions of the MSI in $fMoW-rgb$, as they are effectively downsampled versions of the pan-sharpened RGB images.

4.1. fMoW Splits

We have made the following splits to the dataset:

- `train` Contains 83,412 (62.85%) of the total unique bounding boxes.
- `val` Contains 14,241 (10.73%) of the total unique bounding boxes. This set was made representative of `test`, so that validation can be performed.
- `test` Contains 16,948 (12.77%) of the total unique bounding boxes.
- `seq` Contains 18,115 (13.65%) of the total unique bounding boxes. This set was also made representative of `test`, but was not publicly released during the prize challenge centered around this dataset.

Each split was formed by first binning the GSD, number of temporal views per sequence, UTM zone, and off-nadir angle values. After binning these values, temporal

sequences were divided between the different dataset splits while ensuring that the counts for these bins, as well as the distribution of categories per split, were consistent. Singleton sequences, such as those that are the only ones to cover a particular UTM zone, were also evenly distributed between the various splits. The total number of bounding box instances for each category can be seen in Figure 3.

4.2. fMoW Statistics

Variable length sequences of images are provided for each scene in the dataset. Figure 4 shows the distribution of sequence lengths in fMoW. 21.2% of the sequences contain only 1 view. Most (95%) of the sequences contain 10 or fewer images.

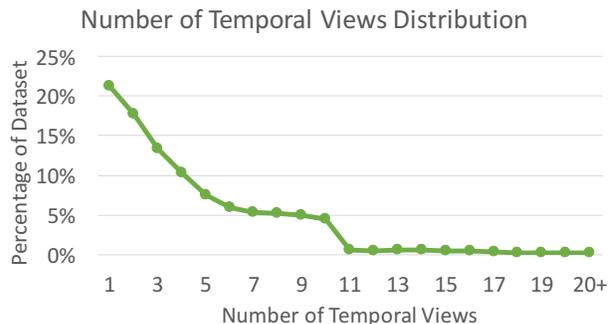


Figure 4: This shows the distribution of the number of temporal views in our dataset. The number of temporal views is not incremented by both the pan-sharpened and multispectral images. These images have almost identical metadata files and are therefore not counted twice. The maximum number of temporal views for any area in the dataset is 41.

A major focus of the collection effort was global diversity. In the metadata, we provide UTM zones, which typically refer to 6° longitude bands (1-60). We also concatenate letters that represent latitude bands (total of 20) to the UTM zones in the metadata. Figure 5 illustrates the frequency of sequences within the UTM zones on earth, where the filled rectangles each represent a different UTM zone. Green colors represent areas with higher numbers of sequences, while blue regions have lower counts. As seen, fMoW covers much of the globe.

The images captured for fMoW also have a wide range of dates, which, in some cases, allows algorithms to analyze areas on earth over long periods of time. Figure 6 shows distributions for years and local times (converted from UTC) in which the images were captured. The average time difference between the earliest and most recent images in each sequence is approximately 3.8 years.

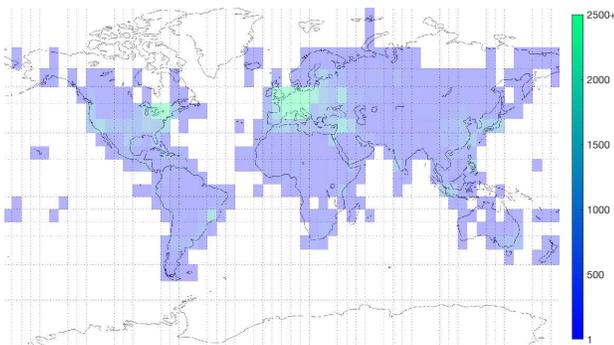


Figure 5: This shows the geographic diversity of fMoW. Data was collected from over 400 unique UTM zones (including latitude bands). This helps illustrate the number of images captured in each UTM zone, where more green colors show UTM zones with a higher number of instances, and more blue colors show UTM zones with lower counts.

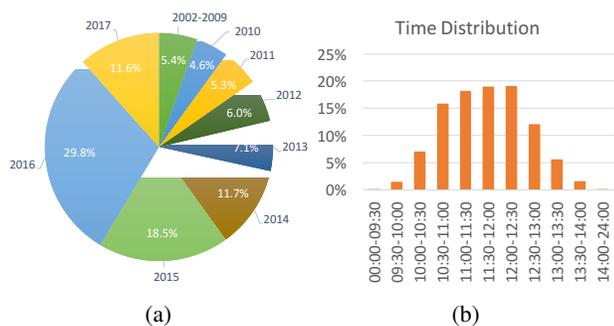


Figure 6: Distribution over (a) years the images were captured, and (b) time of day the images were captured (UTC converted to local time for this figure).

5. Baselines and Methods

Here we present 5 different approaches to our task, which vary by their use of metadata and temporal reason-

ing. All experiments were performed using fMoW-rgb. Two of the methods presented involve fusing metadata into a CNN architecture in an attempt to enable the types of reasoning discussed in the introduction. We perform mean subtraction and normalization for the metadata feature vectors using values calculated over `train + val`.

It is worth noting here that the imagery in fMoW is not registered, and while many sequences have strong spatial correspondence, individual pixel coordinates in different images do not necessarily represent the same positions on the ground. As such, we are prevented from easily using methods that exploit registered sequences.

The CNN used as the base model in our various baseline methods is DenseNet-161 [14], with 48 feature maps ($k=48$). During initial testing, we found this model to outperform other models such as VGG-16 [25] and ResNet-50 [13]. We initialize our base CNN models using the pre-trained ImageNet weights, which we found to improve performance during initial tests. Training is performed using a crop size of 224x224, the Adam optimizer [18], and an initial learning rate of 1e-4. Due to class imbalance in our dataset, we attempted to weight the loss using class frequencies, but did not observe any improvement.

To merge metadata features into the model, the softmax layer of DenseNet is removed and replaced with a concatenation layer to merge DenseNet features with metadata features, followed by two 4096-d fully-connected layers with 50% dropout layers, and a softmax layer with 63 outputs (62 main categories + FD). An illustration of this base model is shown in Figure 7.

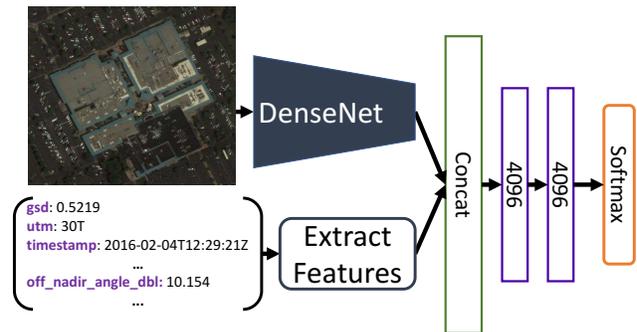


Figure 7: An illustration of our base model used to fuse metadata features into the CNN. This model is used as a baseline and also as a feature extractor (without softmax) for providing features to an LSTM. Dropout layers are added after the 4096-d FC layers.

We test the following approaches with fMoW:

- **LSTM-M** An LSTM architecture trained using temporal sequences of metadata features. We believe training solely on metadata helps understand how important images are in making predictions, while also providing some measure of bias present in fMoW.

- **CNN-I** A standard CNN approach using only images, where DenseNet is fine-tuned after ImageNet. Softmax outputs are summed over each temporal view, after which an argmax is used to make the final prediction. The CNN is trained on all images across all temporal sequences of `train + val`.
- **CNN-IM** A similar approach to CNN-I, but with metadata features concatenated to the features of DenseNet before the fully connected layers.
- **LSTM-I** An LSTM architecture trained using features extracted from CNN-I.
- **LSTM-IM** An LSTM architecture trained using features extracted from CNN-IM.

The LSTM models, which were also trained with the Adam optimizer [18], contained 4096-d hidden states, which were passed to a 512-d multi-layer perceptron (MLP). All of these methods are trained on `train + val`. As tight bounding boxes are typically provided for category instances in the dataset, we add a context buffer around each box before extracting the region of interest from the image. We found that it was useful to provide more context for categories with smaller sizes (*e.g.*, single-unit residential, fountain) and less context for categories that generally cover larger areas (*e.g.*, airports, nuclear power plants).

Per-category F1 scores for `test` are shown in Table 1. From the results, it can be observed that, in general, the LSTM architectures show similar performance to our approaches that sum the probabilities over each view. Some possible contributors to this are the large quantity of single-view images provided in the dataset and that temporal changes may not be particularly important for several of the categories. CNN-I and CNN-IM are also, to some extent, already reasoning about temporal information while making predictions by summing the softmax outputs over each temporal view. Qualitative results that show success and failure cases for LSTM-I are shown in Figure 8. Qualitative results are not shown for the approaches that use metadata, as it is much harder to visually show why the methods succeed in most cases.

It could be argued that the results for approaches using metadata are only making improvements because of bias exploitation. To show that metadata helps beyond inherent bias, we removed all instances from the test set where the metadata-only baseline (LSTM-M) is able to correctly predict the category. The results of this removal, which can be found in Table 2, show that metadata can still be useful for improving performance.

To further confirm the importance of temporal reasoning, we compare the methods presented above with two additional methods, CNN-I-1 and CNN-IM-1, which make predictions for each individual view. We then have all other methods repeat their prediction over the full sequence. This is done to show that, on average, seeing an area multiple

	LSTM-M	CNN-I	LSTM-I	CNN-IM	LSTM-IM
false_detection	0.599	0.728	0.729	0.853	0.837
airport	0.447	0.859	0.800	0.884	0.837
airport hangar	0.017	0.721	0.665	0.677	0.699
airport terminal	0.023	0.697	0.715	0.746	0.759
amusement park	0.622	0.746	0.727	0.898	0.868
aquaculture	0.514	0.754	0.762	0.811	0.805
archaeological site	0.016	0.524	0.491	0.574	0.607
barn	0.292	0.695	0.684	0.717	0.707
border checkpoint	0.000	0.333	0.404	0.523	0.515
burial site	0.019	0.852	0.859	0.827	0.846
car dealership	0.101	0.741	0.797	0.747	0.770
construction site	0.053	0.372	0.373	0.318	0.358
crop field	0.514	0.888	0.872	0.930	0.926
dam	0.158	0.806	0.798	0.864	0.886
debris or rubble	0.381	0.403	0.607	0.474	0.488
educational institution	0.157	0.495	0.475	0.548	0.557
electric substation	0.000	0.849	0.869	0.858	0.872
factory or powerplant	0.000	0.443	0.459	0.536	0.544
fire station	0.028	0.409	0.494	0.483	0.523
flooded road	0.625	0.296	0.285	0.638	0.795
fountain	0.085	0.727	0.705	0.814	0.840
gas station	0.022	0.785	0.779	0.761	0.772
golf course	0.220	0.860	0.916	0.899	0.875
ground transportation station	0.114	0.658	0.694	0.713	0.719
helipad	0.067	0.812	0.856	0.831	0.820
hospital	0.012	0.387	0.404	0.426	0.458
impoverished settlement	0.538	0.410	0.506	0.750	0.704
interchange	0.142	0.833	0.678	0.905	0.909
lake or pond	0.000	0.721	0.650	0.687	0.694
lighthouse	0.037	0.715	0.755	0.779	0.828
military facility	0.426	0.509	0.564	0.597	0.655
multi-unit residential	0.227	0.385	0.414	0.445	0.451
nuclear powerplant	0.000	0.720	0.762	0.600	0.552
office building	0.011	0.198	0.218	0.228	0.225
oil or gas facility	0.522	0.789	0.773	0.844	0.865
park	0.025	0.626	0.638	0.662	0.698
parking lot or garage	0.076	0.775	0.787	0.700	0.732
place of worship	0.362	0.638	0.658	0.712	0.735
police station	0.068	0.246	0.237	0.201	0.329
port	0.444	0.692	0.698	0.736	0.667
prison	0.087	0.611	0.650	0.695	0.726
race track	0.234	0.898	0.886	0.919	0.892
railway bridge	0.030	0.703	0.755	0.761	0.813
recreational facility	0.295	0.907	0.919	0.903	0.906
road bridge	0.000	0.722	0.738	0.747	0.756
runway	0.488	0.821	0.814	0.889	0.885
shipyard	0.000	0.371	0.351	0.368	0.351
shopping mall	0.117	0.615	0.629	0.662	0.662
single-unit residential	0.429	0.688	0.703	0.717	0.684
smokestack	0.204	0.735	0.755	0.772	0.768
solar farm	0.424	0.912	0.921	0.927	0.931
space facility	0.000	0.824	0.737	0.875	0.889
stadium	0.174	0.825	0.850	0.818	0.819
storage tank	0.140	0.921	0.921	0.928	0.924
surface mine	0.200	0.824	0.802	0.870	0.880
swimming pool	0.362	0.920	0.913	0.906	0.907
toll booth	0.030	0.891	0.918	0.960	0.954
tower	0.141	0.723	0.737	0.754	0.777
tunnel opening	0.526	0.867	0.897	0.949	0.942
waste disposal	0.071	0.595	0.570	0.604	0.670
water treatment facility	0.044	0.854	0.816	0.853	0.879
wind farm	0.540	0.939	0.948	0.959	0.968
zoo	0.039	0.566	0.582	0.598	0.611
Average	0.193	0.679	0.688	0.722	0.734

Table 1: F1 scores for different approaches on `test`. Color formatting was applied to each column independently. The average values shown at the bottom of the table are calculated without FD scores.

times outperforms single-view predictions. We note that these tests are clearly not fair for some categories, such as “construction site”, where some views may not even contain the category. However, we perform these tests for completeness to confirm our expectations. Results are shown in Table 3. Per-category results are in the supplement.

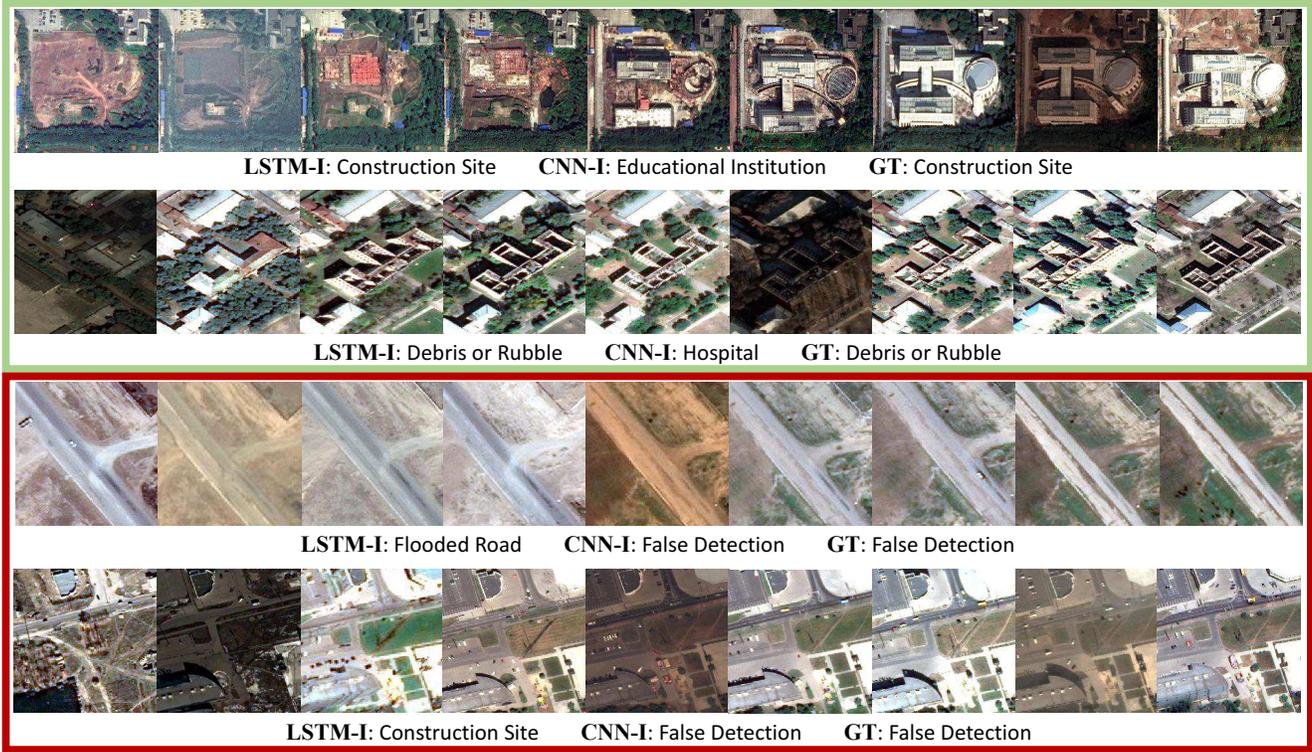


Figure 8: Qualitative examples from `test` of the image-only approaches. The images presented here show the extracted and resized images that are passed to the CNN approaches. The top two rows show success cases for LSTM-I, where CNN-I was not able to correctly predict the category. The bottom two rows show failure cases for LSTM-I, where CNN-I was able to correctly predict the category. Note that sequences with ≥ 9 views were chosen and additional views were trimmed to keep the figure rectangular.

LSTM-M	CNN-I	LSTM-I	CNN-IM	LSTM-IM
0	0.685	0.693	0.695	0.702

Table 2: Results on `test` instances where the metadata-only baseline (LSTM-M) is not able to correctly predict the category. These are the average F1 scores not including FD. These results show that metadata is important beyond exploiting bias in the dataset.

CNN-I-1	CNN-I	LSTM-I	CNN-IM-1	CNN-IM	LSTM-IM
0.618	0.678	0.684	0.666	0.722	0.735

Table 3: Average F1 scores, not including FD, for individual images from `test`. CNN-I-1 and CNN-IM-1 make predictions for each individual view. All other methods repeat their prediction over the full sequence.

6. Conclusion and Discussion

We present fMoW, a dataset that consists of over 1 million satellite images. Temporal views, multispectral im-

agery, and metadata are provided to enable new types of joint reasoning. Models may leverage temporal information and simultaneously reason about the rich set of metadata features (e.g., timestamp, UTM zone) provided for each image. By posing a task in between detection and classification, we avoid the inherent challenges associated with collecting a large geographically-diverse detection dataset, while still allowing for models to be trained that are transferable to real-world detection systems. Different methods were presented for this task that demonstrate the importance of joint reasoning about metadata and temporal information. All code, data, and pretrained models have been made publicly available. We hope that by releasing the dataset and code, other researchers in the CV community will find new and interesting ways to further utilize the metadata and temporal changes to a scene. We also hope to see fMoW being used to train models that are able to assist in humanitarian efforts, such as applications involving disaster relief.

Acknowledgments This work was supported by the Intelligence Advanced Research Projects Activity (IARPA), via Contract 2017-17032700004. Please see our full acknowledgments in the supplement.

References

- [1] ISPRS 2D Semantic Labeling Contest. <http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html>. 3
- [2] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan. YouTube-8M: A Large-Scale Video Classification Benchmark. *arXiv preprint arXiv:1609.08675*, 2016. 2
- [3] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. VQA: Visual Question Answering. In *ICCV*, 2015. 2
- [4] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Nießner, M. Savva, S. Song, A. Zeng, and Y. Zhang. Matterport3D: Learning from RGB-D Data in Indoor Environments. *arXiv preprint arXiv:1709.06158*, 2017. 2
- [5] G. Cheng, J. Han, and X. Lu. Remote Sensing Image Scene Classification: Benchmark and State of the Art. *Proc. IEEE*, 2017. 3
- [6] C. Debes, A. Merentitis, R. Heremans, J. Hahn, N. Frangiadakis, T. van Kasteren, W. Liao, R. Bellens, A. Pižurica, S. Gautama, et al. Hyperspectral and LiDAR Data Fusion: Outcome of the 2013 GRSS Data Fusion Contest. *J-STARs*, 2014. 3
- [7] N. DigitalGlobe, CosmiQ Works. SpaceNet. Dataset available from <https://aws.amazon.com/public-datasets/spacenet/>, 2016. 3
- [8] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term Recurrent Convolutional Networks for Visual Recognition and Description. In *CVPR*, 2015. 2
- [9] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes Challenge: A Retrospective. *IJCV*, 2015. 2
- [10] L. Fei-Fei, R. Fergus, and P. Perona. One-Shot Learning of Object Categories. *PAMI*, 2006. 2
- [11] G. Griffin, A. Holub, and P. Perona. Caltech-256 Object Category Dataset. 2007. 2
- [12] D. Harwath and J. R. Glass. Learning Word-Like Units from Joint Audio-Visual Analysis. *ACL*, 2017. 2
- [13] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2016. 6
- [14] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. Densely Connected Convolutional Networks. *CVPR*, 2017. 6
- [15] N. Jean, M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon. Combining Satellite Imagery and Machine Learning to Predict Poverty. *Science*, 2016. 2
- [16] A. Karpathy and L. Fei-Fei. Deep Visual-Semantic Alignments for Generating Image Descriptions. In *CVPR*, 2015. 2
- [17] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale Video Classification with Convolutional Neural Networks. In *CVPR*, 2014. 2
- [18] D. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. *ICLR*, 2014. 6, 7
- [19] I. Krasin, T. Duerig, N. Alldrin, V. Ferrari, S. Abu-El-Haija, A. Kuznetsova, H. Rom, J. Uijlings, S. Popov, A. Veit, S. Belongie, V. Gomes, A. Gupta, C. Sun, G. Chechik, D. Cai, Z. Feng, D. Narayanan, and K. Murphy. Open-images: A public dataset for large-scale multi-label and multi-class image classification. Dataset available from <https://github.com/openimages>, 2017. 2
- [20] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common Objects in Context. In *ECCV*, 2014. 1, 2
- [21] D. Marmanis, M. Datcu, T. Esch, and U. Stilla. Deep Learning Earth Observation Classification Using ImageNet Pre-trained Networks. *GRSL*, 2016. 2
- [22] C. Padwick, M. Deskevich, F. Pacifici, and S. Smallwood. WorldView-2 Pan-sharpening. In *ASPRS*, 2010. 5
- [23] Y. Pan, T. Mei, T. Yao, H. Li, and Y. Rui. Jointly Modeling Embedding and Translation to Bridge Video and Language. In *CVPR*, 2016. 2
- [24] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 1, 2
- [25] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-scale Image Recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6
- [26] K. Tang, M. Paluri, L. Fei-Fei, R. Fergus, and L. Bourdev. Improving Image Classification with Location Context. In *ICCV*, 2015. 2
- [27] S. Wang, M. Bai, G. Mattyus, H. Chu, W. Luo, B. Yang, J. Liang, J. Cheverie, S. Fidler, and R. Urtasun. Torontocity: Seeing the world with a million eyes. *ICCV*, 2017. 3
- [28] T. Weyand, I. Kostrikov, and J. Philbin. Planet-photo geolocation with convolutional neural networks. In *ECCV*, 2016. 2
- [29] G.-S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, and X. Lu. AID: A Benchmark Data Set for Performance Evaluation of Aerial Scene Classification. *TGRS*, 2017. 3
- [30] G.-S. Xia, X.-Y. Tong, F. Hu, Y. Zhong, M. Datcu, and L. Zhang. Exploiting Deep Features for Remote Sensing Image Retrieval: A Systematic Investigation. *arXiv preprint arXiv:1707.07321*, 2017. 2
- [31] Y. Yang and S. Newsam. Bag-Of-Visual-Words and Spatial Extensions for Land-Use Classification. In *ACM GIS*, 2010. 3
- [32] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond Short Snippets: Deep Networks for Video Classification. In *CVPR*, 2015. 2