Satlas: A Large-Scale, Multi-Task Dataset for Remote Sensing Image Understanding

Favyen Bastani  Piper Wolters  Ritwik Gupta  Joe Ferdinando  Aniruddha Kembhavi
PRIOR @ Allen Institute for AI

Figure 1. SATLAS is a large scale dataset for remote sensing. Labels in this dataset are relevant to many important earth monitoring applications, including water resource monitoring, tracking deforestation, detecting wind turbines for mapping renewable energy infrastructure, extracting ice extents for tracking glacier loss, flood detection, tracking urban expansion, and vessel detection for tackling illegal fishing.

Abstract

Remote sensing images are useful for a wide variety of environmental and earth monitoring tasks, including tracking deforestation, illegal fishing, urban expansion, and natural disasters. The earth is extremely diverse—the amount of potential tasks in remote sensing images is massive, and the sizes of features range from several kilometers to just tens of centimeters. However, creating generalizable computer vision methods is a challenge in part due to the lack of a large-scale dataset that captures these diverse features for many tasks. In this paper, we present SATLAS, a remote sensing dataset and benchmark that is large in both breadth, featuring all of the aforementioned applications and more, as well as scale, comprising 290M labels under 137 categories and seven label modalities. We evaluate eight baselines and a proposed method on SATLAS, and find that there is substantial room for improvement in addressing research challenges specific to remote sensing, including processing image time series that consist of images from very different types of sensors, and taking advantage of long-range spatial context. We also find that pre-training on SATLAS substantially improves performance on downstream tasks with few labeled examples, increasing average accuracy by 16% over ImageNet and 5% over the next best baseline.

1. Introduction

Satellite and aerial images provide a diverse range of information about the physical world. In images of urban areas, we can identify unmapped roads and buildings and incorporate them into digital map datasets, as well as monitor urban expansion. In images of industrial areas, we can catalogue solar farms and wind turbines to track the progress of renewable energy deployment. In images of glaciers and forests, we can monitor slow natural changes like glacier loss and deforestation. With the availability of global, regularly updated, and public domain sources of remote sensing images like the EU’s Sentinel missions [3], we can monitor the earth for all of these applications at global-scale on a monthly or even weekly basis.

Because the immense scale of the earth makes global manual analysis of remote sensing images cost-prohibitive, automatic computer vision methods are crucial for unlocking their full potential. Previous work has proposed applying computer vision for automatically inferring the positions of roads and buildings [9, 12, 30, 34, 56, 57]; moni-
toring changes in land cover and land use such as deforestation and urban expansion [43, 44]; predicting vessel positions and types to help tackle illegal fishing [40]; and tracking the progress and extent of natural disasters like floods, wildfires, and tornadoes [6, 22, 42]. However, in practice, most deployed applications continue to rely on manual or semi-automated rather than fully automated analysis of remote sensing images [1] for two reasons. First, accuracy remains a barrier even in major applications like road extraction [11], making full automation impractical. Second, there is a long tail of remote sensing applications that require expert annotation but have few labeled examples (e.g., a recent New York Times study manually documented illegal airstrips in Brazil using satellite images [7]).

We believe that the lack of a very-large-scale, multi-task remote sensing dataset is a major impediment for progress on automated methods for remote sensing tasks today. State-of-the-art architectures such as Vision Transformers [24] and CLIP [41] require huge datasets to achieve peak performance. However, existing remote sensing datasets for object detection, instance segmentation, and semantic segmentation like DOTA [51], iSAID [54], and DeepGlobe [23] contain less than 10K images each, compared to the 328K images in COCO and millions used to train CLIP; the small size of these datasets means we cannot fully take advantage of recent architectures. Additionally, existing remote sensing benchmarks are fragmented, with individual benchmarks for categories like roads [39], vessels [40], and crop types [26], but no benchmark spanning many categories. The lack of a large-scale, centralized, and accessible benchmark prevents transfer learning opportunities across tasks and modalities, and makes it difficult for computer vision researchers to engage in this domain.

In this paper, we present SATLAS, a large-scale multi-task dataset for benchmarking and improving remote sensing image understanding models. Our goal with SATLAS is to label everything that is visible in a satellite image. To this end, SATLAS includes 290M distinct labels under 137 categories and seven diverse label modalities: these are points like wind turbines and water towers; polygons like buildings and airports; polylines like roads and rivers; segmentation and regression labels like land cover categories and bathymetry values (water depth); properties of objects like the rotor diameter of a wind turbine; and patch classification labels like the presence of smoke in an image. Figure 1 demonstrates the wide range of categories in SATLAS, along with the diverse applications that they serve.

We believe that SATLAS will encourage work on computer vision methods that tackle the unique research challenges in the remote sensing domain. Compared to general-purpose computer vision methods, remote sensing models require specialized techniques such as accounting for long-range spatial context (detecting an airport taxiway requires both high-resolution imagery to detect the taxiway, but also potentially several km of context around the taxiway to understand its relation to the airport) and synthesizing information across several images over time (classifying the type of crop grown at a crop field requires observations at several stages of the agricultural cycle [26]). Furthermore, remote sensing images are captured by diverse sensor types, including multispectral bands like short wave infrared and non-optical sensors like synthetic aperture radar, and images vary substantially in ground sample distance resolution (SATLAS includes medium-resolution 10 m/pixel images and high-resolution 1 m/pixel images). Finally, objects of interest vary widely in size, from forests spanning many km² to street lamps smaller than one square meter.

We evaluate eight computer vision baselines on SATLAS and find that no single existing method is able to accommodate all of the diverse label modalities in SATLAS; instead, each baseline can only predict a subset of categories. Thus, inspired by recent work that integrate task-specific output heads [20, 27, 32, 33], we develop a unified model called SATNET that incorporates three such heads so that it can learn from every category in the dataset. SATNET also incorporates a simple but effective feature pooling step to learn from image time series. When jointly trained on all categories and then fine-tuned for each label modality, SATNET is able to leverage transfer learning opportunities between the modalities to improve average performance.

In addition to evaluating on SATLAS, we evaluate the effectiveness of pre-training on SATLAS for fine-tuning on seven downstream tasks, compared to pre-training on other datasets as well as self-supervised learning methods. We find that SATLAS pre-training significantly improves performance over the baselines, showing that SATLAS can readily improve accuracy on the numerous niche remote sensing tasks that require costly expert annotation.

In summary, our contributions are: (1) SATLAS, a large-scale remote sensing benchmark with 137 categories under seven diverse label modalities, (2) SATNET, a unified model that supports predictions for all label modalities, and (3) Demonstrating that pre-training on SATLAS improves downstream performance by 5% on average.

We will release the dataset at https://satlas.allenai.org.

2. Related Work

Large-Scale Remote Sensing Datasets. Several general-purpose remote sensing computer vision datasets have been released. Many of these focus on scene and patch classification tasks: the UC Merced Land Use (UCM) [53] and BigEarthNet [48] datasets involve land cover classification with 21 and 43 categories respectively, while the AID [52], Million-AID [37], RESISC45 [18], and Functional Map of the World (FMoW) [21] datasets additionally include
Self-supervised Learning for Remote Sensing.

3. Satlas

We present SATLAS, a very-large-scale dataset and benchmark for remote sensing that improves on existing datasets in three key ways:

1. Scale: SATLAS contains 10x more image pixels than the largest existing dataset, FMoW, while covering 50x more of the earth.

2. Label diversity and multi-modality: Existing datasets in Table 1 are uni-modal in their labels, focusing e.g. on only classification or only object detection. SATLAS labels span seven modalities; furthermore, they comprise 137 categories, 2x more than the largest existing dataset.

3. Spatio-temporal images and labels: rather than being tied to individual remote sensing images, our labels are associated with geographic coordinates (i.e., longitude-latitude positions) and time ranges. This enables methods to make predictions from multiple images across time, as well as leverage long-range spatial context from neighboring images. These features present new research challenges that, if solved, can greatly improve model performance.

3.1. Label Modalities and Categories

SATLAS labels span 137 categories, with seven key distinct modalities that we summarize in Figure 2:

1. Semantic segmentation—e.g., predicting per-pixel land cover (water vs forest vs developed vs etc.).
2. Regression—e.g., predicting per-pixel bathymetry (water depth) or percent tree cover.
3. Points (object detection)—e.g., predicting wind turbines, oil wells, and vessels.
4. Polygons (instance segmentation)—e.g., predicting buildings, dams, and aquafarms.
5. Polylines—e.g., predicting roads, rivers, and railways.
6. Properties of points, polygons, and polylines—e.g., the rotor diameter of a wind turbine.
7. Classification labels—e.g., whether an image exhibits negligible, medium, or high wildfire smoke density.

Incorporating these diverse modalities makes the dataset more challenging for individual models to solve, while also enabling SATLAS to serve as a more general-purpose remote sensing benchmark. We derive labels from a combination of new annotation by workers on Amazon Mechanical Turk and processing existing data sources like maps.

Segmentation. We incorporate five segmentation tasks in SATLAS that together comprise 36 categories: land cover, current water levels, crop type, burned areas and fire retardant, and floods and clouds.

First, land cover and land use maps describe a combination of the earth’s surface material (e.g., trees, water, asphalt, or ice) and how the land is used by humans (e.g.,
developed areas versus uninhabited areas). We process the WorldCover map [50] to derive 11 land cover and land use categories, ranging from barren land to developed areas. Tracking changes in land cover has numerous applications, such as monitoring urban expansion, monitoring deforestation, and identifying growing and receding water bodies.

Second, we annotate 10K images for the current coastline contour visible in each image. While land cover maps include a category for water bodies, they focus on a long time horizon, labeling water based on the mean high water level (i.e., the average level of water at high tide). The actual water level can vary greatly both through the tide cycle and over seasonal changes. Current water levels are of interest to earth and environmental scientists, while the contours themselves can improve accuracy for detecting coastal features like near-shore vessels and platforms.

Third, we derive crop type labels from OpenStreetMap, a collaborative, openly licensed map dataset built through edits made by contributing users. In 20K images, we process 16 categories including soy, rice, and corn. Crop type mapping is useful for both environmental monitoring and food price prediction.

Fourth, we annotate 10K images for two wildfire-related tasks: fire retardant drops and burned areas. Automated methods for these tasks can help inform both real-time fire response efforts, and longer-term forest management practices. Fire retardant is a substance dropped along the boundaries of a wildfire to contain or slow its progression. We define burned areas as grass or tree covered areas recently burned within the last three months of the current image.

Fifth, we include flood and cloud labels from C2S [4].

Regression. SATLAS includes two regression tasks that involve predicting numerical values at each pixel.

First, we process coastal bathymetry lidar scans from NOAA to derive bathymetry (water depth) data in SATLAS. Prior work has shown that water depth (bathymetry) up to around 10 m can be measured from optical satellite images [8]. Coastal bathymetry has several applications, including near-shore ship navigation and for monitoring changes in water reservoir volumes.

Second, we annotate tree cover canopy percentage in 3K images. Tree cover data can capture forest growth and loss on a more fine-grained scale than land cover maps.

Points, Polygons, and Polylines. We derive labels for 49 categories of points, polygons, and polylines from OpenStreetMap, including points like oil wells, lighthouses, and water towers; polygons like buildings, parking lots, and wastewater plants; and polylines like roads, railways, and airport taxiways and runways. Automatically detecting roads and buildings that are missing from digital maps like OpenStreetMap can reduce the cost of improving those maps. Other features like wind turbines, solar farms, and crop fields are useful for tracking renewable energy deployment progress and agricultural forecasting.

We annotate three additional categories. Two are short-term categories that do not appear in maps: airplanes and vessels. We represent airplanes as points, labeling the point at which the wings connect to the fuselage. Likewise, we annotate the center point of vessels. Vessel detection in satellite images has recently attracted interest because it is useful for tackling illegal fishing [40]: for example, systems can alert authorities if they detect vessels with visible fishing equipment in marine protected areas.

The third longer-term category is rooftop solar panels, which we also represent as points. Rooftop panel installations are not accurately mapped in existing datasets, in part because of their recent rapid growth. That same growth makes monitoring the rate of new installations useful to researchers studying renewable energy deployment.

Properties. We derive labels for 12 properties from Open-
StreetMap. Five are numerical: the rotor diameter and electricity output of a wind turbine, the capacity of a parking lot, and the number of lanes and maximum speed of a road. The other seven are categorical, and span 35 categories: the type of a sports track (e.g., running, cycling), the type of a road (e.g., motorway, residential street), the type of a park (e.g., city park, sports pitch), the type of a power plant (e.g., oil, nuclear), the material produced by a quarry (e.g., sand, gravel), and the presence of a bridge on a road (binary).

Classification. We also annotate images for two three-category scene classification tasks indicating the degree of snow accumulation (none, some, and substantial) and the density of smoke or haze (none, partial, and substantial).

Figure 3 shows a frequency curve of categories in SATLAS. Figure 4 enumerates all 137 categories.

3.2. Imagery

Image Types. SATLAS includes a repository of 5 million megapixels of remote sensing images spanning 85 million km² of the earth’s surface. We include both medium-resolution (10 m/pixel) images captured by Sentinel-2 that cover much of the earth, and high-resolution (1 m/pixel) images captured under the US National Agriculture Imagery Program (NAIP) available in the continental US. These substantial differences in resolution make SATLAS more realistic, since often both high-resolution but outdated images and recent, medium-resolution images are available; they also make SATLAS more challenging, since methods must learn to leverage both types of images.

Figure 5 shows examples of both NAIP and Sentinel-2 images, and highlights the diversity of areas covered by SATLAS. Figure 6 shows the global geographic coverage.

Associating Images with Labels. Existing datasets (including all but FMoW in Table 1) typically associate each label with a single image, and require methods to predict the label with that one image only. In reality, though, image providers regularly release newly captured images: for example, Sentinel-2 provides images of a given location roughly once a week. Accuracy often improves significantly when leveraging a time series of several images.

Rather than associate labels directly to individual images, we annotate each label with geographic (longitude-latitude) coordinates and an estimated validity period specifying the time range where the label is physically present; this may be a few seconds for an airplane or many years for a building. We also expose the geographic extent and capture time of each image, and allow methods to use all of the images in the dataset in arbitrary ways. This way, methods can learn to synthesize features across several images to improve prediction accuracy; for example, when predicting the crop type grown at a crop field, observations of the crop field at different stages of the agricultural cycle can provide different clues about the type of crop grown there. Additionally, this enables methods to make use of long-range spatial context from many neighboring images; for example, when observing an airport terminal building, understanding the relationship of the building within the airport is needed to identify it as a terminal.

Splits. We split the SATLAS images into three splits: training (80% of the tiles), validation (10%), and test (10%). We will host a leaderboard for the test split.

Because SATLAS includes diverse Sentinel-2 and NAIP images, we define two separate evaluation modes: in the full mode, methods can make use of both Sentinel-2 and NAIP images, while in the medium-resolution mode, methods only input the Sentinel-2 images. For the medium-resolution mode, we exclude categories that are too small to accurately observe at the 10 m/pixel resolution.

4. SatNet

Off-the-shelf computer vision model architectures are not able to predict the wide range of label modalities in SATLAS: for example, while Mask2Former [17] can simultaneously perform semantic and instance segmentation, it is not designed to predict properties of polygons or classify image patches. This prevents these models from taking advantage of the full set of transfer learning opportunities that are present; for example, detecting building polygons is likely useful for segmenting images for land cover and land use, since those tasks include a human-developed category. We develop a unified model, SatNet, that is capable of learning from all label modalities in SATLAS.

Figure 7 shows a schematic of SatNet. SatNet is inspired by recent work that employ task-specific output heads [20, 27, 32], as well as methods that synthesize features across remote sensing image time series [21, 25]. It has the following design. We extend a Swin Transformer [36] backbone with three task-specific output heads: we use the standard Swin head for classification, a UNet-like head [45] for segmentation and regression, and Mask R-CNN [28] in conjunction with a Feature Pyramid Network [35] for detecting points and polygons. For polylines, while special-
ized polylines extraction architectures have been shown to improve accuracy [9, 30, 49], we opt to employ the simpler segmentation approach [56] where we apply the UNet head to segment images for polylines categories, and post-process the segmentation probabilities with binary thresholding, morphological thinning, and line following and simplification [19] to extract polylines.

In SatNet, we also incorporate a simple but effective approach for synthesizing features across multiple images: when an image time series is available at a location, we process each image in the series through the Swin backbone, aggregate features across the images using max pooling, and then apply the task-specific heads on the aggregated features. We also tested Conv-LSTMs and video transformers, but found that the simple max feature pooling approach performed comparably in preliminary experiments; nevertheless, we believe there is substantial room for improvement in better leveraging remote sensing image time series.

5. Evaluation

We first evaluate our method and eight classification, semantic segmentation, and instance segmentation baselines on the SATLAS test split in Section 5.1. We then evaluate performance on seven downstream applications in Section 5.2, comparing pre-training on SATLAS to pre-training on other remote sensing datasets, as well as self-supervised learning techniques specialized for remote sensing.

5.1. Results on Satlas

Methods. We compare SatNet against eight baselines on SATLAS. We select baselines that are either standard models or models that provide state-of-the-art performance for subsets of task types in SATLAS. None of the baselines are able to handle the full range of SATLAS label modalities. For scene and property classification and regression, we compare ResNet [29], Vision Transformer [24], and Swin Transformer [36]. For segmentation, regression,
Table 2. Results on the SATLAS test set when using all images and when using only the medium-resolution Sentinel-2 images. We break down results by seven label modalities: segmentation (Seg), regression (Reg), points (Pt), polygons (Pgon), polylines (Pline), properties (Prop), and classification (Cls). We show absolute error for Reg (lower is better), and accuracy for the other modalities (higher is better).

| Method | All Images | Only Medium-Resolution Sentinel-2 Images |
|--------|------------|----------------------------------------|
|        | Seg ▼ Reg ▼ Pt ▼ Pgon ▼ Pline ▼ Prop ▼ Cls ▼ | Seg ▼ Reg ▼ Pt ▼ Pgon ▼ Pline ▼ Prop ▼ Cls ▼ |
| PSPNet (ResNext-101) [35] | 0.77 | 20 | - | - | 0.34 | - | - | 0.44 | 14 | - | 0.19 | - | - |
| LinkNet (ResNext-101) [14] | 0.64 | 13 | - | - | 0.21 | - | - | 0.36 | 12 | - | 0.13 | - | - |
| DeepLabv3 (ResNext-101) [15] | 0.65 | 11 | - | - | 0.31 | - | - | 0.50 | 12 | - | 0.22 | - | - |
| ResNet-50 [29] | - | - | - | - | 0.70 | 0.96 | - | - | - | - | 0.49 | 0.96 | - | - |
| VIT-Large [24] | - | - | - | - | 0.80 | 0.99 | - | - | - | - | 0.49 | 0.99 | - | - |
| Swin-Base [36] | - | - | - | - | 0.80 | 0.98 | - | - | - | - | 0.50 | 0.98 | - | - |
| Mask R-CNN (Swin-Base) [28] | - | - | 0.32 | 0.21 | - | - | - | - | 0.16 | 0.10 | - | - | - | - |
| ISTR [31] | - | - | 0.02 | 0.05 | - | - | - | - | 0.01 | 0.01 | - | - | - | - |
| SatNet (single-image, per-modality) | 0.76 | 6 | 0.38 | 0.22 | 0.37 | 0.87 | 0.99 | 0.47 | 7 | 0.15 | 0.10 | 0.17 | 0.53 | 0.99 |
| SatNet (single-image, joint training) | 0.76 | 8 | 0.33 | 0.18 | 0.37 | 0.90 | 0.99 | 0.42 | 11 | 0.08 | 0.09 | 0.20 | 0.64 | 0.99 |
| SatNet (single-image, fine-tuning) | 0.78 | 6 | 0.37 | 0.20 | 0.45 | 0.91 | 0.99 | 0.53 | 7 | 0.16 | 0.11 | 0.28 | 0.64 | 0.99 |
| SatNet (multi-image, fine-tuning) | 0.78 | 6 | 0.38 | 0.20 | 0.45 | 0.91 | 0.99 | 0.55 | 7 | 0.17 | 0.11 | 0.33 | 0.64 | 0.99 |

Figure 5. SATLAS incorporates both high-resolution NAIP and medium-resolution Sentinel-2 images that exhibit very different characteristics. In (c) and (d), comparing multiple images helps distinguish a rye crop field and a fire retardant drop.

Figure 6. Top: geographic coverage of SATLAS, with bright pixels indicating locations covered by images and labels in the dataset. SATLAS spans all continents except Antarctica. Bottom: temporal coverage around Europe and Africa.

Across all methods, we employ random cropping, horizontal and vertical flipping, and random resizing augmentations during training. We initialize models with ImageNet-pretrained weights. We use the Adam optimizer, and initialize the learning rate to $10^{-4}$, decaying via halving down to $10^{-6}$ upon plateaus in the training loss. We weight each training example based on the maximum inverse frequency of categories appearing in the example, so that rare categories are sampled with a balanced frequency. We train with a batch size of 32 for 50K batches, which takes 50 hours on 4 NVIDIA RTX A6000 GPUs.

**Metrics.** We use standard metrics for each modality: accuracy for scene and property classification, F1 score for semantic segmentation, mean absolute error for regression, mAP accuracy for point detection and instance segmentation, and GEO accuracy [13] for polyline extraction.

**Results.** We show results on Satlas in Table 2. We report performance for four variants of SATNET:

- **SatNet (per-modality):** train separately per-modality.
- **SatNet (joint training):** jointly train across all categories and modalities.
- **SatNet (fine-tuning):** fine-tune the jointly trained parameters on each modality.
- **SatNet (multi-image, fine-tuning):** leverage the full image time series at each location (the other variants, like the baselines, use the most recent image only).
Across the seven label modalities, SATNET is able to match the performance or surpass the state-of-the-art, purpose-built baseline methods when trained separately on each modality, validating SATNET’s effectiveness as a unified model that can be applied for a diverse range of remote sensing tasks. Jointly training one set of SATNET parameters for all categories reduces average performance on several modalities, but SATNET remains competitive in most cases; this training mode provides large efficiency gains since the backbone features need only be computed once for each image during inference, rather than once per modality. When fine-tuning SATNET on each modality using the parameters derived from joint training, SATNET provides accuracy improvements for segmentation, regression, and property prediction tasks. This supports our hypothesis that there are transfer learning opportunities across the modalities, validating the utility of a unified model for improving performance. SATNET with image time series inputs provides improved performance for some modalities, such as segmentation and polyline tasks with Sentinel-2 images only, but the improvements are slight; we believe that improved methods for synthesizing features across multiple images have the potential to greatly improve accuracy.

We show qualitative results in Figure 8. We achieve high accuracy on several categories, such as wind turbines and water towers. However, for oil wells, one well is detected but several others are not. Similarly, for polyline features like roads and railways, the model produces short segments and other noisy outputs, despite ample training data for these categories; we believe that incorporating and improving models that are tailored for specialized output types like polylines [30, 49] has the potential to improve accuracy.

5.2. Downstream Performance

We now evaluate accuracy on seven downstream applications when pre-training on SATLAS compared to pre-training on four existing remote sensing datasets, as well as two self-supervised learning methods. For each downstream application, we evaluate few-shot accuracy when training on different percentages of the dataset, to focus on the challenge of improving performance on niche remote sensing applications that require expert annotation and thus have few labeled examples.

**Methods.** We compare pre-training on SATLAS to pre-training on four existing remote sensing datasets: BigEarthNet [48], Million-AID [37], DOTA [51], and iSAID [54]. We use SATNET in all cases, restoring the pre-trained Swin backbone and then fine-tuning on each downstream dataset.

We also compare two self-supervised learning methods, Momentum Contrast v2 (MoCo) [16] and Seasonal Contrast (SeCo) [38]. The latter is a specialized method for remote sensing that leverages multiple image captures of the same location to learn invariance to seasonal changes. For MoCo, we use our SATNET model and train on SATLAS images. For SeCo, we evaluate three variants: their original model with their dataset, their model with SATLAS, and SATNET with SATLAS. We fine-tune the weights learned through self-supervision on the downstream tasks.

We fine-tune both the pre-training and self-supervised learning methods by first freezing the backbone and only training the prediction head for 32K examples, and then fine-tuning the entire model. After unfreezing the backbone, we increase the learning rate linearly from 0 to $10^{-4}$ over another 32K examples.

**Downstream Datasets.** We measure performance on seven downstream tasks. Four are existing large-scale remote sensing datasets that involve classification with between 21 and 63 categories: UCM [53], AID [52], RESISC45 [18], and FMoW [21]. The other three are the Massachusetts Buildings and Massachusetts Roads datasets [39], which involve semantic segmentation, and the Airbus Ships [2] dataset, which involves instance segmentation.

**Results.** Figure 9 shows downstream performance with varying training set sizes. SATLAS consistently outperforms the baselines: when training on 50 examples, we improve average accuracy across the tasks by 5% over the next best baseline, and 16% over ImageNet pre-training.

6. Releasing Model Predictions

Many SATLAS labels are directly useful for earth, environmental, and social science applications. We plan to col-
Figure 8. Qualitative results on the SATLAS test set. The rightmost image shows a failure case where SatNet detects only one of 5 oil wells.

Figure 9. Results on downstream tasks, with varying number of training examples. SATLAS pre-training consistently improves downstream accuracy, showing it can readily improve performance on the numerous remote sensing tasks that have few labeled examples.

select models trained on SATLAS that have sufficient accuracy on specific applications like mapping renewable energy deployment and tracking damage from natural disasters, and compute their outputs globally on a regular basis as new satellite imagery becomes available. We intend to release these predictions on a centralized platform where non-ML researchers can download the data for free and apply it in their work. As progress is made on remote sensing research challenges in the computer vision community, and better models become available that enable new applications, we plan to integrate these new models into the platform.

We will also release the Satlas dataset and code at https://satlas.allenai.org.

7. Conclusion

By improving on existing datasets in both scale and label diversity, SATLAS serves as an effective very-large-scale benchmark for remote sensing methods. We believe that this will encourage further work on tackling the remaining research challenges in this domain, especially in leveraging image time series and long-range spatial context. We have also shown that SATLAS substantially improves downstream accuracy, which is crucial for the long tail of remote sensing tasks that have few labeled examples. Finally, we intend to release global predictions made by models trained on SATLAS that are directly useful for earth, environmental, and social science researchers.

References

[1] The machine vision challenge to better analyze satellite images of Earth. MIT Technology Review. 2
[2] Airbus ship detection challenge. https://www.kaggle.com/c/airbus-ship-detection, 2018. Airbus. 8
[3] Copernicus Sentinel Missions. https://sentinel.esa.int/web/sentinel/home, 2022. European Space Agency. 1
[4] A global flood events and cloud cover dataset (version 1.0), 2022. Cloud to Street, Microsoft, Radiant Earth Foundation. 4
[5] Peri Akiva, Matthew Purri, and Matthew Leotta. Self-supervised Material and Texture Representation Learning for Remote Sensing Tasks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 8203–8215, 2022. 3
[6] Robert S Allison, Joshua M Johnston, Gregory Craig, and Sion Jennings. Airborne Optical and Thermal Remote
Fernando Paolo, Tsu ting Tim Lin, Ritwik Gupta, Bryce Ronghang Hu and Amanpreet Singh. Unit: Multimodal Multitask Learning with a Unified Transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 1439–1449, 2021. 2, 5

Amrita Kamath, Christopher Clark, Tanmay Gupta, Eric Kolve, Derek Hoiem, and Aniruddha Kembhavi. Webly Supervised Concept Expansion for General Purpose Vision Models. arXiv preprint arXiv:2202.02317, 2022. 2

Zuo Yue Li, Jan Dirk Wegner, and Aurélien Lucchi. Topological Map Extraction from Overhead Images. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 1715–1724, 2019. 1

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Bharath Hariharan, and Serge Belongie. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 2117–2125, 2017. 5

Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In IEEE/CVF International Conference on Computer Vision (ICCV), pages 9992–10002, 2021. 5, 6, 7

Yang Long, Gui-Song Xia, Shengyang Li, Wen Yang, Michael Ying Yang, Xiao Xiang Zhu, Liangpei Zhang, and Deren Li. On Creating Benchmark Dataset for Aerial Image Interpretation: Reviews, Guidelines, and Million-AID. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, pages 4205–4230, 2021. 2, 3, 8

Oscar Manas, Alexandre Lacoste, Xavier Giro i Nieto, David Vazquez, and Pau Rodriguez. Seasonal Contrast: Unsupervised Pre-Training from Uncurated Remote Sensing Data. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021. 3, 8

Volodymyr Mnih. Machine Learning for Aerial Image Labeling. PhD thesis, University of Toronto, 2013. 2, 8

Fernando Paolo, Tsu ting Tim Lin, Ritwik Gupta, Bryce Goodman, Nirav Patel, Daniel Kuster, David Kroodsma, and Jared Dumnnon. xView-3-SAR: Detecting Dark Fishing Activity Using Synthetic Aperture Radar Imagery. Neural Information Processing Systems, 2022. 2, 3, 4

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning Transferable Visual Models from Natural Language Supervision. International Conference on Machine Learning, pages 8748–8763, 2021. 2

Sudha Radhika, Yukio Tamura, and Masahiro Matsui. Application of Remote Sensing Images for Natural Disaster Mitigation using Wavelet based Pattern Recognition Analysis. In 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pages 84–87, 2016. 2

Caleb Robinson, Le Hou, Kolya Malkin, Rachel Soobitsky, Jacob Czawlytko, Bistra Dilikina, and Nebojsa Jojic. Large Scale High-Resolution Land Cover Mapping with Multi-Resolution Data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 12726–12735, 2019. 2

Caleb Robinson, Anthony Ortiz, Kolya Malkin, Blake Elias, Andi Peng, Dan Morris, Bistra Dilikina, and Nebojsa Jojic. Human-Machine Collaboration for Fast Land Cover Mapping. pages 2509–2517, 2020. 2

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 234–241. Springer, 2015. 5

Linus Scheibenreif, Joëlle Hanna, Michael Momment, and Damian Borth. Self-Supervised Vision Transformers for Land-Cover Segmentation and Classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 1422–1431, 2022. 3

Linus Scheibenreif, Michael Momment, and Damian Borth. Contrastive Self-Supervised Data Fusion for Satellite Imagery. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 3:705–711, 2022. 3

Gencer Sumbul, Marcela Charfuen, Begum Demir, and Volker Markl. BigEarthNet: A Large-Scale Benchmark Archive for Remote Sensing Image Understanding. In International Geoscience and Remote Sensing Symposium (IGARSS), 2019. 2, 3, 8

Yong-Qiang Tan, Shang-Hua Gao, Xuan-Yi Li, Ming-Ming Cheng, and Bo Ren. VecRoad: Point-based Iterative Graph Exploration for Road Graphs Extraction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 8910–8918, 2020. 6, 8

Ruben Van De Kerchove, Daniele Zanaga, Wanda Keersmaecker, Niels Souverijns, Jan Wevers, Carsten Brockmann, Alex Grous, Audrey Paccini, Oliver Cartus, Maurizio Santoro, et al. ESA WorldCover: Global land cover mapping at 10 m resolution for 2020 based on Sentinel-1 and 2 data. In AGU Fall Meeting Abstracts, volume 2021, pages GC45I–GC451, 2021. 4

Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Be longie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. DOTA: A Large-scale Dataset for Object Detection in Aerial Images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018. 2, 3, 8

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. IEEE Journal of Transactions on Geoscience and Remote Sensing, 55(7):3965–3981, 2017. 2, 3, 8

Yi Yang and Shawn Newsam. Bag-Of-Visual-Words and Spatial Extensions for Land-Use Classification. ACM Conference on Spatial Information (SIGSPATIAL), 2010. 2, 3, 8

Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Khan, Guolei Sun, Fahad Shahbaz Khan, Fan Zhu, Ling Shao, Gui-Song Xia, and Xiang Bai. iSAD: A Large-scale Dataset for Instance Segmentation in Aerial Images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2019. 2, 3, 8

Hengshuang Zhao, Jianping Shi, Xiaojian Qi, Xiaogang Wang, and Jiaya Jia. Pyramid Scene Parsing Network. In...
[56] Lichen Zhou, Chuang Zhang, and Ming Wu. D-LinkNet: LinkNet with Pretrained Encoder and Dilated Convolution for High Resolution Satellite Imagery Road Extraction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pages 182–186, 2018.

[57] Stefano Zorzi, Shabab Bazrafi, Stefan Habenschuss, and Friedrich Fraundorfer. PolyWorld: Polygonal Building Extraction with Graph Neural Networks in Satellite Images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 1848–1857, 2022.