DeepGlobe 2018: A Challenge to Parse the Earth through Satellite Images

İlke Demir[1], Krzysztof Koperski[2], David Lindenbaum[3], Guan Pang[1],
Jing Huang[1], Saikat Basu[1], Forest Hughes[1], Devis Tuia[4], Ramesh Raskar[5]

[1]Facebook, [2]DigitalGlobe, [3]CosmiQ Works,
[4]Wageningen University, [5] The MIT Media Lab

Figure 1: **DeepGlobe Challenges:** Example road extraction, building detection, and land cover classification training images superimposed on corresponding satellite images.

Abstract

We present the DeepGlobe 2018 Satellite Image Understanding Challenge, which includes three public competitions for segmentation, detection, and classification tasks on satellite images (Figure 1). Similar to other challenges in computer vision domain such as DAVIS[21] and COCO[33], DeepGlobe proposes three datasets and corresponding evaluation methodologies, coherently bundled in three competitions with a dedicated workshop co-located with CVPR 2018.

We observed that satellite imagery is a rich and structured source of information, yet it is less investigated than everyday images by computer vision researchers. However, bridging modern computer vision with remote sensing data analysis could have critical impact to the way we understand our environment and lead to major breakthroughs in global urban planning or climate change research. Keeping such bridging objective in mind, DeepGlobe aims to bring together researchers from different domains to raise awareness of remote sensing in the computer vision community and vice-versa. We aim to improve and evaluate state-of-the-art satellite image understanding approaches, which can hopefully serve as reference benchmarks for future research in the same topic. In this paper, we analyze characteristics of each dataset, define the evaluation criteria of the competitions, and provide baselines for each task.

1. Introduction

As machine learning methods dominate the computer vision field, public datasets and benchmarks have started to play an important role for relative scalability and reliability of different approaches. Driven by community efforts such as ImageNet [45] for object detection, COCO [33] for image captioning, and DAVIS [21] for object segmentation, computer vision research had been able to push the limits of what we can achieve, by using the same annotated datasets and common training/validation conventions. Such datasets and corresponding challenges increase the visibility, availability, and feasibility of machine learning models, which brought up even more scalable, diverse, and accurate algorithms to be evaluated on public benchmarks.

We observe that satellite imagery is a powerful source of information as it contains more structured and uniform data compared to traditional images. Although computer vision community has been accomplishing hard tasks on everyday image datasets using deep learning and in contrast to public datasets released for everyday media, satellite images are only recently gaining attention from the community for map composition, population analysis, effective precision agriculture, and autonomous driving tasks.

To direct more attention to such approaches, we present DeepGlobe 2018, a Satellite Image Understanding Challenge, which (i) contains three datasets structured to solve three different satellite image understanding tasks, (ii) organizes three public challenges to propose solutions to these tasks, and (iii) gathers researchers from diverse fields to unite all expertises to solve similar tasks in a collaborative workshop. The datasets created and released for this competition may serve as (iv) fair and durable reference benchmarks for future research in satellite image analysis. Furthermore, since the challenge tasks involve "in the wild" forms of classic computer vision problems (e.g., image classification, detection, and semantic segmentation), these datasets have the potential to become valuable testbeds for the design of robust vision algorithms, beyond the area of remote sensing.

The three tracks for DeepGlobe are defined as follows:

- **Road Extraction Challenge:** In disaster zones, especially in developing countries, maps and accessibility information are crucial for crisis response. We pose the challenge of automatically extracting roads and street networks remotely from satellite images as a first step for automated crisis response and increased map coverage for connectivity.

- **Building Detection Challenge:** As evidenced by recent catastrophic natural events, modeling population dynamics is of great importance for disaster response and recovery. Thus, modeling urban demographics is a vital task and detection of buildings and urban areas are key to achieve it. We pose the challenge of automatically detecting buildings from satellite images for gathering aggregate urban information remotely as well as for gathering detailed information about spatial distribution of urban settlements.

- **Land Cover Classification Challenge:** Automatic categorization and segmentation of land cover is essential for sustainable development, agriculture [11], forestry [17, 16] and urban planning [20]. Therefore, we pose the challenge of classifying land types from satellite images for economic and agricultural automation solutions, among the three topics of DeepGlobe, probably as the most studied one in the intersection of remote sensing and image processing.

We currently host three public competitions based on the tasks of extracting roads, detecting buildings, and classifying land cover types in the satellite images. The combined datasets include over 10K satellite images. Section 2 explains the characteristics of images, details the annotation process, and introduces the division of training, validation, and test sets. Section 3 describes the tasks in detail and proposes the evaluation metric used for each task. Section 4 provides an overview of the current approaches and gives our preliminary baselines for the competitions.

The results of the competitions will be presented in the DeepGlobe 2018 Workshop during the 2018 International Conference on Computer Vision and Pattern Recognition (CVPR) in Salt Lake City, Utah on June 18th, 2018. As of May 15st, 2018, more than 950 participants have registered in DeepGlobe competitions and there are more than 90 valid submissions in the leaderboard over the three tracks.

2. Datasets

In this Section, we will discuss the dataset and imagery characteristics for each DeepGlobe track, followed by an explanation of the methodology for the annotation process to obtain the training labels.

2.1. Road Extraction

There have been several datasets proposed in the literature for benchmarking algorithms for semantic segmentation of overhead imagery. Some of these can be enumerated as the TorontoCity[54] dataset, the ISPRS 2D semantic labeling dataset [3], the Mnih dataset [39], the SpaceNet dataset [2] and the ISPRS Benchmark for Multi-Platform Photogrammetry [4].

The satellite imagery used in DeepGlobe for the road extraction challenge is sampled from the DigitalGlobe +Vivid Images dataset [1]. It covers images captured over Thailand, Indonesia, and India. The ground resolution of the image pixels is 50 cm/pixel. The images consist of 3 channels (Red, Green and Blue). Each of the original geotiff images are $19'584 \times 19'584$ pixels. The annotation process starts by tiling and loading these images in QGIS[7]. Based on this tiling, we determine useful areas to sample from those countries. For designating useful areas, we sample data uniformly between rural and urban areas. After sampling we select the corresponding DigitalGlobe tiff images belonging to those areas. These images are then cropped to extract useful subregions and relevant subregions are sampled by GIS experts. (A useful subregion denotes a part of the image where we have a good relative ratio between positive and negative examples.) Also, while selecting these subregions, we try to sample interesting areas uniformly, e.g., those with different types of road surfaces (unpaved, paved, dirt roads), rural and urban areas, etc. An example of one image crop is illustrated in the left panel of Figure 1. It is

Figure 2: Road labels are annotated on top of the satellite image patches, all taken from DeepGlobe Road Extraction Challenge dataset.

important to note that the labels generated are pixel-based, where all pixels belonging to the road are labeled, instead of labeling only the centerline.

The final road dataset consists of a total of $8'570$ images and spans a total land area of $2'220km^2$. Of those, $6'226$ images (72.7% of the dataset), spanning a total of $1'632km^2$, were split as the training dataset. $1'243$ images, spanning $362km^2$, were chosen as the validation dataset and $1'101$ images were chosen for testing which cover a total land area of $288km^2$. The split of the dataset to training/validation/testing subsets is conducted by randomizing among tiles to aim for an approximate distribution of 70%/15%/15%. The training dataset consists of 4.5% positive and 95.5% negative pixels, the validation dataset consists of 3% positive and 97% negative pixels and the test dataset consists of 4.1% positive and 95.9% negative pixels. We selected a diverse set of patches to demonstrate road labels annotated on the original satellite images in Figure 2. As shown, the urban morphology, the illumination conditions, the road density, and the structure of the street networks are significantly diverse among the samples.

2.2. Building Detection

DeepGlobe Building Detection Challenge uses the *SpaceNet* Building Detection Dataset. Previous competitions on building extraction using satellite data, PRRS 2008 [10] and ISPRS [3, 5], were based on small areas (a few km^2) and in some cases used a combination of opti-

cal data and LiDAR data. The Inria Aerial Image Labeling covered $810km^2$ area with 30cm resolution in various European and American cities [35]. It addressed model portability between areas as some cities were included only in training data and some only in testing data. SpaceNet was the first challenge that involved large areas including cities in Asia and Africa.

The dataset includes four areas: Las Vegas, Paris, Shanghai, and Khartoum. The labeled dataset consists of $24'586$ $200m \times 200m$ (corresponding to 650×650 pixels) non-overlapping scenes containing a total of $302'701$ building footprints across all areas. The areas are of urban and suburban nature. The source imagery is from the WorldView-3 sensor, which has both a $31cm$ single-band panchromatic and a $1.24m$ multi-spectral sensor providing 8-band multi-spectral imagery with 11-bit radiometric resolution. A GIS team at DigitalGlobe (now Radiant Solutions) fully annotated each scene, identifying and providing a polygon footprint for each building to the published specification, which were extracted to best represent the building footprint (see the central panel of Figure 1 for an example). Any partially visible rooftops were approximated to represent the shape of the building. Adjoining buildings were marked as a single building. The dataset was split 60%/20%/20% for train/validation/test. As per the nature of human-based building annotation, some small errors are inevitable especially for rural areas. We leave the analysis of annotater disagreement for future work.

Each area is covered by a single satellite image, which constitutes an easier problem to solve compared to data where different parts are covered by images having different sun and satellite angles, and different atmospheric conditions. Atmospheric compensation process can process images to create data that reflects surface reflectance therefore reducing effects of atmosphere, but different shadow lengths and different satellite orientation can possibly create problems for detection algorithms if models are used to classify images acquired at different time with different sun/satellite angles.

The *SpaceNet* data[9] is distributed under a Creative Commons Attribution-ShareAlike 4.0 International License and is hosted as a public dataset on Amazon Web Services and can be downloaded for free.

2.3. Land Cover Classification

Semantic segmentation started to attract more research activities as a challenging task. The ISPRS Vaihingen and Potsdam [3] and the Zeebruges data [22] are popular public datasets for this task. The ISPRS Vaihingen dataset contains 33 images of different size (on average $2'000 \times 2'000$ pixels), with 16 fully annotated images. ISPRS Potsdam contains 38 images of size $6'000 \times 6'000$ pixels, with 24 annotated images. Annotations for both datasets have 6

classes. Vaihingen and Potsdam are both focused in urban city area, with classes limited to urban targets such as buildings, trees, cars. Zeebruges is a 7-tiles dataset (each one of size $10'000 \times 10'000$) with 8 classes (both land cover and objects), acquired both by RGB images at 5cm resolution and a LiDAR point cloud). Dstl Kaggle dataset [47] covered $1km^2$ of urban area with RGB and 16-band (including SWIR) WorldView-3 images. Besides urban areas, another important application of land cover classification is humanitarian studies focusing more on rural areas at mid-resolution ($\sim 30m/pixel$). For similar problems Landsat data is (i.e., crop type classification[31]). Still, the low resolution of Landsat data limits the information it can provide.

The DeepGlobe Land Cover Classification Challenge is the first public dataset offering high-resolution sub-meter satellite imagery focusing on rural areas. Due to the variety of land cover types and to the density of annotations, this dataset is more challenging than existing counterparts described above. DeepGlobe Land Cover Classification Challenge dataset contains $1'146$ satellite images of size $2'448 \times 2'448$ pixels in total, split into training/validation/test sets, each with 803/171/172 images (corresponding to a split of 70%/15%/15%). All images contain RGB data, with a pixel resolution of 50 cm, collected from the DigitalGlobe Vivid+ dataset as described in Section 2.1. The total area size of the dataset is equivalent to $1'716.9km^2$.

Each satellite image is paired with a mask image for land cover annotation. The mask is an RGB image with 7 classes following the Anderson Classification [14]. The class distributions are available in Table 1.

- Urban land: Man-made, built up areas with human artifacts.

- Agriculture land: Farms, any planned (i.e. regular) plantation, cropland, orchards, vineyards, nurseries, and ornamental horticultural areas; confined feeding operations.
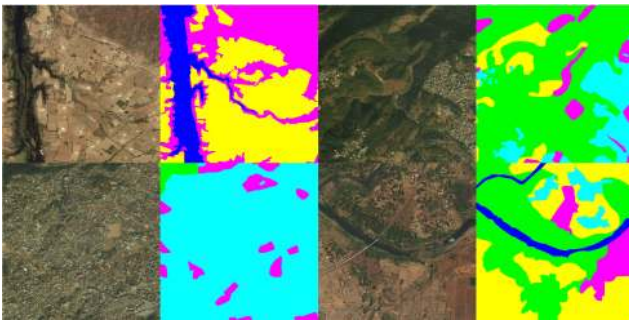

Figure 3: Some example land cover class label (right) and corresponding original image (left) pairs from interesting areas. Label colors are given in Table 1.

- Rangeland: Any non-forest, non-farm, green land, grass.

- Forest land: Any land with at least 20% tree crown density plus clear cuts.

- Water: Rivers, oceans, lakes, wetland, ponds.

- Barren land: Mountain, rock, dessert, beach, land with no vegetation.

- Unknown: Clouds and others.

class	pixel count	proportion
Urban	642.4M	9.35%
Agriculture	3898.0M	56.76%
Rangeland	701.1M	10.21%
Forest	944.4M	13.75%
Water	256.9M	3.74%
Barren	421.8M	6.14%
Unknown	3.0M	0.04%

Table 1: Class distributions in the DeepGlobe land cover classification dataset.

The annotations are pixel-wise segmentation masks created by professional annotators (see the right hand panel in Figure 1). The images in the dataset were sampled from full-size tiles to assure that all land cover classes have enough representation. In our specifications, we indicated that any instance of a class larger than a roughly $20m \times 20m$ would be annotated. However, land cover pixel-wise classification from high-resolution satellite imagery is still an exploratory task, and some small human error is inevitable. In addition, we intentionally did not annotate roads and bridges because it is already covered in the road extraction challenge. Some example labeled areas are demonstrated in Figure 3 as examples of farm, forest, and urban dominant tiles, and a mixed tile.

3. Tasks and Evaluations

In this section, we formally define the expected task in each challenge and explain the evaluation metrics used in terms of their computation and implementation.

3.1. Road Extraction

We formulate the task of road extraction from satellite images as a binary classification problem. Each input is a satellite image. The solution is expected to predict a mask for the input (i.e., a binary image of the same height and width as the input with road and non-road pixel labels).

There have been previous challenges on road mask extraction, e.g., the SpaceNet [2]. Their metric was based on

the Averaged Path Length Similarity (APLS) metric [51] that measures distance between ground truth road network represented in vector form with a solution graph also in vector form. Any proposed road graph G with missing edges (e.g., if an overhanging tree is inferred to sever a road) is penalized by the APLS metric, so ensuring that roads are properly connected is crucial for a high score.

In our case, we use the pixel-wise Intersection over Union (IoU) score as our evaluation metric for each image, defined as Eqn. (1).

$$IoU_i = \frac{TP_i}{TP_i + FP_i + FN_i},$$ (1)

where TP_i is the number of pixels that are correctly predicted as road pixel, FP_i is the number of pixels that are wrongly predicted as road pixel, and FN_i is the number of pixels that are wrongly predicted as non-road pixel for image i. Assuming there are n images, the final score is defined as the average IoU among all images (Eqn. (2)).

$$mIoU = \frac{1}{n} \sum_{i=1}^{n} IoU_i$$ (2)

3.2. Building Detection

In DeepGlobe, building detection is based on a binary segmentation task, where the input is a satellite image and the output is a list of building polygons. Multiple performance measures can be applied to score participants. PRRS 2008 [10] challenge used 8 different performance measures. Our evaluation metric for this competition is an F1 score with the matching algorithm inspired by Algorithm 2 in the ILSVRC paper applied to the detection of building footprints [45]. This metric was decided to emphasize the importance of both accurate detection of buildings and the importance of complete identification of building outlines in an area. Buildings with a pixel area of 20 pixels or less were discarded, as these small buildings are artifacts of the image tiling process when a tile boundary cuts a building into multiple parts.

A detected building is scored as a true positive if the IoU (Eqn.3) between the ground truth (GT) building area A and the detected building area B is larger than 0.5. If a proposed building intersects with multiple GT buildings, then the GT building with the largest IoU value will be selected.

$$IoU = \frac{Area\,(A \cap B)}{Area\,(A \cup B)}$$ (3)

The solution score is defined by F1 measure (Eqn. 4), where TP is number of true positives, M is the number of ground truth buildings and N is the number of detected buildings.

$$F1 = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} = \frac{2 * TP}{M + N}$$ (4)

The implementation and detailed description of scoring can be found in the SpaceNet Utilities GitHub repo [9]. We score each area separately and the final score is the average of scores for each area as in Eqn. 5.

$$F1 = \frac{F1_{AOI1} + F1_{AOI2} + F1_{AOI3} + F1_{AOI4}}{4}$$ (5)

3.3. Land Cover Classification

The land cover classification problem is defined as a multi-class segmentation task to detect areas of classes mentioned in Section 2.3. The evaluation is computed based on the accuracy of the class labels and averages over classes are considered. The class 'unknown' is removed from the evaluation, as it does not correspond to a land cover class, but rather to the presence of clouds.

Each input is a satellite image. The solution is expected to predict an RGB mask for the input, i.e., a colored image of the same height and width as the input image. The expected result is a land cover map of same size in pixels as the input image, where the color of each pixel indicates its class label.

There have been previous challenges on road mask extraction (e.g., the TiSeLaC [8]), which emphasized the usage of temporal information of the dataset. Our challenge, on the other hand, uses images captured at one timestamp as the input, thus more flexible in real applications. Other previous land cover / land use semantic segmentation challenges as the ISPRS [3] or the IEEE GRSS data fusion contests [22, 57] also used single shot ground truths and reported overall and average accuracy scores as evaluation metrics.

Similar to the road extraction challenge, we use the pixel-wise Intersection over Union (IoU) score as our evaluation metric. It was defined slightly differently for each class, as there are multiple categories (Eqn. 6). Assuming there are n images, the formulation is defined as,

$$IoU_j = \frac{\sum_{i=1}^{n} TP_{ij}}{\sum_{i=1}^{n} TP_{ij} + \sum_{i=1}^{n} FP_{ij} + \sum_{i=1}^{n} FN_{ij}},$$ (6)

where TP_{ij} is the number of pixels in image i that are correctly predicted as class j, FP_{ij} is the number of pixels in image i that are wrongly predicted as class j, and FN_{ij} is the number of pixels in image i that are wrongly predicted as any class other than class j. Note that we have an unknown class that is not active in our evaluation (i.e., the predictions on such pixels will not be added to the calculation and thus do not affect the final score). Assuming there are k land cover classes, the final score is defined as the average IoU among all classes as in Eqn. (7).

$$mIoU = \frac{1}{k} \sum_{j=1}^{k} IoU_j$$ (7)

4. State-of-the-art and Baselines

The tasks defined in the previous section have been explored in different datasets with different methods, some of which are also shared publicly. In this section we will introduce the state of the art approaches for each task comparing their dataset to DeepGlobe. As a baseline, we will also share our preliminary results based on current approaches on each dataset, which sets the expected success figures for the challenge participants as well as guide them to develop novel approaches.

4.1. Road Extraction

Automating the generation of road networks is extensively studied in the computer vision and computer graphics world. The procedural generation of streets [13, 23] creates detailed and structurally realistic models, however written grammars are not based on the real world. On the other hand, some inverse procedural modeling (IPM) approaches [12] process real-world data (images, LiDAR, etc.) to extract realistic representations. Following the example-based generation idea, another approach is to use already existing data resources, such as aerial images [37, 38], or geostationary satellite images [26, 58]. Similar approaches extract road networks using neural networks for dynamic environments [53] from LiDAR data [59], using line integrals [32] and using image processing approaches [43, 55].

Similar to the experiments of [26] and [37], we explored our baseline approach to follow some state-of-the-art deep learning models [19, 24, 28, 44]. In contrast to those approaches, our dataset is more diverse, spanning three countries with significant changes in topology and climate; and significantly larger in area and size. The best results were obtained by training a modified version of DeepLab [24] architecture with ResNet18 backbone and Focal Loss [49]. In order to provide a baseline to evaluate the network, we only added simple rotation as data augmentation, and we did not apply any post-processing to the results, only binarizing all results at a fixed threshold of 128. With this setup, we obtained an IoU score of 0.545 after training 100 epochs. Two example results are given in Figure 4, showing the satellite image, extracted road mask, and ground truth road mask from left to right. The vanishing roads suggest that post-processing techniques other than simple thresholding would yield more connected roads.

4.2. Building Detection

Building detection and building footprint segmentation has been subject of research for long time. Early work was based on pan-sharpened images and was using land cover classification to find vegetation, shadows, water and man-made areas followed by segmentation and classifica-

Figure 4: Example results of our road extraction method with an IoU score of 0.545, with satellite image (left), extracted road mask (middle), and the ground truth (right).

tion of segments into building/non-building areas [10]. Researchers sometimes transformed pixels into HSV color space, which alleviates effects of different illumination on pitched roofs. In [41] the author used shadow information and vegetation/shadow/man-made classification combined with graph approach to detect buildings.

Mnih [40] used two locally connected NN layers followed by a fully connected layer. He also took into account omission noise (some objects are not marked in the ground truth data) by modifying loss function and mis-registration noise (such noise exists if the ground truth is not based on image, but on some other data, such as OSM [6] or survey data) by allowing for translational noise. Vakalopoulou et al. [50] used convolutional layers of AlexNet to extract features that were used as input to SVM classifier which was classifying pixels into building/non-building classes. Saito and Aoki [46] used CNN based approaches for building and road extraction. Liu et al. [34] used FCN-8 segmentation network analyzing IR, R and G data with 5 convolutional layers and augmentation with a model based on nDSM (normalized Digital Surface Model) and NDVI. Inria competition solutions described in [29] used U-Net or SegNet approaches to segmentation.

The current approach to building detection on our dataset uses the top scoring solutions from the SpaceNet Building Detection Round 2 result. The final results from the 2017 competition are shown in Table 2. It is important to note that top algorithms performed best in Las Vegas and worst in Khartoum, the visible structural organization and illumination variance in different urban morphologies are probable causes for this performance loss in the Khartoum data. The winning algorithm by competitor XD_XD used an ensemble of 3 U-Net models [44] to segment an 8-band multi-spectral image with additional use of OpenStreetMap [6] data and then to extract building footprints from the segmentation

Rank	Competitor	Total Score	Individual City Scores			
			Las Vegas	Paris	Shanghai	Khartoum
1	XD_XD	0.693	0.885	0.745	0.597	0.544
2	wleite	0.643	0.829	0.679	0.581	0.483
3	nofto	0.579	0.787	0.584	0.520	0.424

Table 2: Final results from SpaceNet Building Detection Challenge Round2 (2017), as a baseline for building detection.

output. An ensemble classifier was trained on each of the 4 AOIs individually . This segmentation approach produced high scores for IoU with an average larger than 0.8, while the IoU threshold for the competition is 0.5. The algorithm struggles with small objects and in locations where buildings are very close to each other. The detailed descriptions of the algorithms can be found in [25, 42]. Figure 5 shows the high performance of the algorithm in Las Vegas and in the bottom left you can see the algorithm has problems extracting close buildings in Paris.

Building detection can be followed by building footprint extraction, which can be used with DSM information to create 3D models of buildings [5, 15, 27]. 3D models can be combined with material classification and images taken from oblique angles to create accurate and realistic models for large scenes [30].

4.3. Land Cover Classification

Land Cover Classification from satellite imagery is a very active research problem in remote sensing. Earlier work mostly focus on image classification, where each image is only classified to one label. Yang and Newsam[56] used Bag-of-Visual-Words and SVM to classify a dataset of $2'100$ images containing 21 classes each with 100 images of size 256×256 pixels. The best accuracy reported[56] was 81.19%, and the data was released as UC Merced dataset[56], which later became a popular dataset for land cover image classification. Scott et al.[48] applied DCNN-based approach on the UCM dataset and obtained a best accuracy of 98.5% with ResNet-50.

Compared to image-level classification problem, pixel-wise classification, or semantic segmentation, started to attract more research activities as a challenging task, for which deep learning pipelines are becoming the state of the art [60]. Volpi and Tuia [52] proposed to use an downsample-upsample CNN-based architecture and obtained an F1 score of 83.58% on Vaihingen and 87.97% on Potsdam. Audebert et al. [18] trained a variant of the SegNet architecture with multi-kernel convolutional layer, and achieved 89.8% on Vaihingen. Marmanis and colleagues [36] achieved similar performances using an ensemble of CNNs models trained to recognized classes and using edges information between classes. Authors in [22] performed a comparison of several state of art algorithms on

Figure 5: Results from SpaceNet baseline. From top left, clockwise: Vegas, Vegas, Khartoum, Paris. In all panels, the blue outline represents the ground truth, the green outlines are true positives, the red are false positives.

the Zeebruges dataset, including kernel-based, fine-tuned VGG models and CNN trained from scratch.

We also performed pixel-wise classification on our DeepGlobe land cover data, by designing a CNN architecture based on DeepLab [24] using ResNet18 backbone with atrous spatial pyramid pooling (ASPP) block and batch normalization. We used data augmentation by integrating rotations and also weighted classes based on class distributions (see Table 1). This approach achieved an IoU score of 0.433 at epoch 30 with a 512×512 patch size.

Example results are demonstrated in Figure 6, our result on the left, satellite image in the middle, and ground truth on the right. Note that the results and the IoU scores reported are the direct segmentation results from our model without post-processing. Harder distinctions like farms, rangelands, and forests are well-conceived by our model (third and last rows). Small structures not annotated in the ground truth, such as little ponds (top-left), and narrow beaches (second
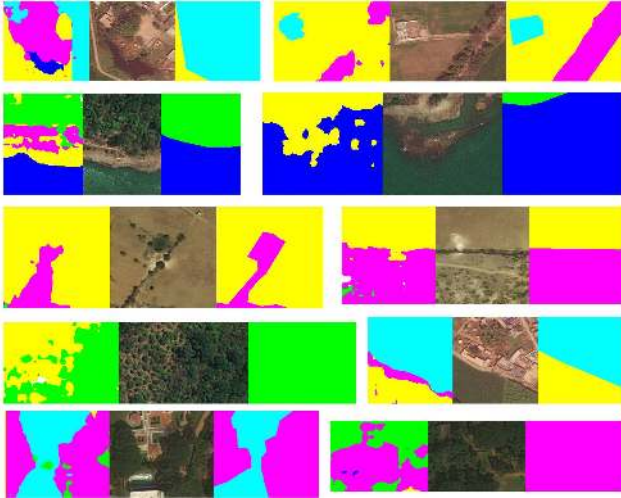
Figure 6: Some result patches (left) produced by our land cover classification baseline approach, paired with the corresponding satellite image (middle) and the ground truth (right).

row left and fourth row right) are also correctly classified by our model. Such cases, however, decreases the IoU score. Although the granularity of the segmentation looks superior to the ground truth (left of first two rows), applying a CRF or a clustering approach would improve the IoU scores.

5. Conclusions

We introduced the DeepGlobe Satellite Image Understanding Challenge. DeepGlobe 2018 provides datasets, challenges, and a CVPR 2018 workshop structured around the three tasks of understanding roads, buildings, and land cover types from satellite images. In this paper, we analyzed the datasets and explained the evaluation metrics for the public challenges. We also provided some baselines compared to the state-of-the-art approaches. Based on the current feedback, we believe that the DeepGlobe datasets will become valuable benchmarks in satellite image understanding, enabling more collaborative interdisciplinary research in the area, that can be fairly compared and contrasted using our benchmark, leading to new exciting developments at the intersection of computer vision, machine learning, remote sensing, and geosciences.

Acknowledgments

References

[1] DigitalGlobe Basemap +Vivid. `https://dg-cms-uploads-production.s3.amazonaws.com/uploads/document/file/2/DG_Basemap_Vivid_DS_1.pdf`.

[2] Introducing the SpaceNet road detection and routing challenge and dataset. `https://medium.com/the-downlinq/introducing-the-spacenet-road-detection-and-routing-challenge-and-dataset-7604de39b779`.

[3] ISPRS 2d semantic labeling dataset. `http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html`.

[4] ISPRS benchmark for multi-platform photogrammetry. `http://www2.isprs.org/commissions/comm1/icwg15b/benchmark_main.html`.

[5] ISPRS test project on urban classification and 3d building reconstruction. `http://www2.isprs.org/commissions/comm3/wg4/detection-and-reconstruction.html`.

[6] OpenStreetMap. `openstreetmap.org`.

[7] QGIS. `https://qgis.org/en/site/`.

[8] TiSeLaC: Time series land cover classification challenge. `https://sites.google.com/site/dinoienco/tiselc`.

[9] SpaceNet on Amazon Web Services (AWS). "datasets." the SpaceNet catalog. `https://spacenetchallenge.github.io/datasets/datasetHomePage.html`, Last modified April 30, 2018.

[10] S. Aksoy, B. Ozdemir, S. Eckert, F. Kayitakire, M. Pesarasi, O. Aytekin, C. C. Borel, J. Cech, E. Christophe, S. Duzgun, A. Erener, K. Ertugay, E. Hussain, J. Inglada, S. Lefevre, O. Ok, D. K. San, R. Sara, J. Shan, J. Soman, I. Ulusoy, and R. Witz. Performance evaluation of building detection and digital surface model extraction algorithms: Outcomes of the prrs 2008 algorithm performance contest. In *2008 IAPR Workshop on Pattern Recognition in Remote Sensing (PRRS 2008)*, pages 1–12, Dec 2008.

[11] C. Alcantara, T. Kuemmerle, A. V. Prishchepov, and V. C. Radeloff. Mapping abandoned agriculture with multi-temporal MODIS satellite data. *Remote Sens. Enviro.*, 124:334–347, 2012.

[12] D. G. Aliaga, I. Demir, B. Benes, and M. Wand. Inverse procedural modeling of 3d models for virtual worlds. In *ACM SIGGRAPH 2016 Courses*, SIGGRAPH '16, pages 16:1–16:316, New York, NY, USA, 2016. ACM.

[13] D. G. Aliaga, C. A. Vanegas, and B. Benes. Interactive example-based urban layout synthesis. *ACM Trans. Graph.*, 27(5):160:1–160:10, Dec. 2008.

[14] J. R. Anderson, E. E. Hardy, J. T. Roach, and R. E. Witmer. A land use and land cover classification system for use with remote sensor data. Technical report, 1976.

[15] H. Arefi and P. Reinartz. Building reconstruction using dsm and orthorectified images. *Remote Sensing*, 5(4):1681–1703, 2013.

[16] G. P. Asner, E. N. Broadbent, P. J. C. Oliveira, M. Keller, D. E. Knapp, and J. N. M. Silva. Condition and fate of logged forests in the Brazilian Amazon. *Proc. Nat. Ac. Science (PNAS)*, 103(34):12947–12950, 2006.

[17] G. P. Asner, D. Knapp, E. Broadbent, P. Oliveira, M. Keller, and J. Silva. Ecology: Selective logging in the Brazilian Amazon. *Science*, 310:480–482, 2005.

[18] N. Audebert, B. L. Saux, and S. Lefevre. Semantic segmentation of earth observation data using multimodal and multi-scale deep networks. *Asian Conference on Computer Vision*, 2016.

[19] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *CoRR*, abs/1511.00561, 2015.

[20] B. Bechtel, M. Demuzere, Y. Xu, M.-L. Verdonck, P. Lopes, L. See, C. Ren, F. V. Coillie, D. Tuia, C. C. Fonte, A. Cassone, N. Kaloustian, O. Conrad, M. Tamminga, and G. Mills. Beyond the urban mask: Local climate zones as a generic descriptor of urban areas. potential and recent developments. In *Joint Urban Remote Sensing Event*, Dubai, UAE, 2017.

[21] S. Caelles, A. Montes, K.-K. Maninis, Y. Chen, L. Van Gool, F. Perazzi, and J. Pont-Tuset. The 2018 DAVIS Challenge on Video Object Segmentation. *ArXiv e-prints*, Mar. 2018.

[22] M. Campos-Taberner, A. Romero-Soriano, C. Gatta, G. Camps-Valls, A. Lagrange, B. L. Saux, A. Beaupère, A. Boulch, A. Chan-Hon-Tong, S. Herbin, H. Randrianarivo, M. Ferecatu, M. Shimoni, G. Moser, and D. Tuia. Processing of extremely high resolution LiDAR and RGB data: Outcome of the 2015 IEEE GRSS Data Fusion Contest. Part A: 2D contest. *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, 9(12):5547–5559, 2016.

[23] G. Chen, G. Esch, P. Wonka, P. Müller, and E. Zhang. Interactive procedural street modeling. *ACM Trans. Graph.*, 27(3):103:1–103:10, Aug. 2008.

[24] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99):1–1, 2017.

[25] L. David. 2nd SpaceNet competition winners code release. https://medium.com/the-downlinq/ 2nd-spacenet-competition-winners-code-release-c7473eea7c11.

[26] I. Demir, F. Hughes, A. Raj, K. Dhruv, S. M. Muddala, S. Garg, B. Doo, and R. Raskar. Generative street addresses from satellite imagery. *ISPRS International Journal of Geo-Information*, 7(3), 2018.

[27] I. Garcia-Dorado, I. Demir, and D. G. Aliaga. Automatic urban modeling using volumetric reconstruction with surface graph cuts. *Computers & Graphics*, 37(7):896–910, Nov. 2013.

[28] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016.

[29] B. Huang, K. Lu, N. Audebert, A. Khalel, Y. Tarabalka, J. Malof, A. Boulch, B. Le Saux, L. Collins, K. Bradbury, et al. Large-scale semantic classification: outcome of the first year of inria aerial image labeling benchmark. In *IEEE International Geoscience and Remote Sensing Symposium–IGARSS 2018*, 2018.

[30] IARPA. Creation of operationally realistic 3D environment (CORE3D). https://www.iarpa.gov/index.php/research-programs/core3d.

[31] N. Kussul, M. Lavreniuk, S. Skakun, and A. Shelestov. Deep learning classification of land cover and crop types using remote sensing data. *IEEE Geoscience and Remote Sensing Letters*, 2017.

[32] P. Li, Y. Zang, C. Wang, J. Li, M. Cheng, L. Luo, and Y. Yu. Road network extraction via deep learning and line integral convolution. In *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 1599–1602, July 2016.

[33] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*. European Conference on Computer Vision, September 2014.

[34] Y. Liu, S. Piramanayagam, S. T. Monteiro, and E. Saber. Dense semantic labeling of very-high-resolution aerial imagery and lidar with fully convolutional neural networks and higher-order crfs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Honolulu, USA*, 2017.

[35] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In *IEEE International Symposium on Geoscience and Remote Sensing (IGARSS)*, 2017.

[36] D. Marmanis, K. Schindler, J. Wegner, S. Galliani, M. Datcu, and U. Stilla. Classification with an edge: Improving semantic image segmentation with boundary detection. *ISPRS Journal of Photogrammetry and Remote Sensing*, 135:158 – 172, 2018.

[37] G. Mattyus, W. Luo, and R. Urtasun. Deeproadmapper: Extracting road topology from aerial images. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[38] G. Mattyus, S. Wang, S. Fidler, and R. Urtasun. Enhancing road maps by parsing aerial images around the world.

In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1689–1697, Dec 2015.

[39] V. Mnih. *Machine Learning for Aerial Image Labeling*. PhD thesis, University of Toronto, 2013.

[40] V. Mnih and G. E. Hinton. Learning to label aerial images from noisy data. In *Proceedings of the 29th International conference on machine learning (ICML-12)*, pages 567–574, 2012.

[41] A. O. Ok. Automated detection of buildings from single vhr multispectral images using shadow information and graph cuts. *ISPRS journal of photogrammetry and remote sensing*, 86:21–40, 2013.

[42] K. Ozaki. Winning solution for the SpaceNet challenge: Joint learning with OpenStreetMap. `https://i.ho.lc/winning-solution-for-the-spacenet-challenge-joint-learning-with-openstreetmap.html`.

[43] R. Peteri, J. Celle, and T. Ranchin. Detection and extraction of road networks from high resolution satellite images. In *Proceedings 2003 International Conference on Image Processing (Cat. No.03CH37429)*, volume 1, pages I–301–4 vol.1, Sept 2003.

[44] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351 of *LNCS*, pages 234–241. Springer, 2015. (available on arXiv:1505.04597 [cs.CV]).

[45] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[46] S. Saito and Y. Aoki. Building and road detection from large aerial imagery. In *Image Processing: Machine Vision Applications VIII*, volume 9405, page 94050K. International Society for Optics and Photonics, 2015.

[47] D. Science and T. L. (Dstl). Dstl satellite imagery feature detection. `https://www.kaggle.com/c/dstl-satellite-imagery-feature-detection`.

[48] G. Scott, M. England, W. Starms, R. Marcum, and C. Davis. Training deep convolutional neural networks for land cover classification of high-resolution imagery. *IEEE Geoscience and Remote Sensing Letters*, 2017.

[49] R. G. K. H. P. D. Tsung-Yi Lin, Priya Goyal. Focal loss for dense object detection. In *ICCV*, 2017.

[50] M. Vakalopoulou, K. Karantzalos, N. Komodakis, and N. Paragios. Building detection in very high resolution multispectral data with deep learning features. In *Geoscience and Remote Sensing Symposium (IGARSS), 2015 IEEE International*, pages 1873–1876. IEEE, 2015.

[51] A. Van Etten. SpaceNet road detection and routing challenge - part I. `https://medium.com/the-downlinq/spacenet-road-detection-and-routing-challenge-part-i-d4f59d55bfce`.

[52] M. Volpi and D. Tuia. Dense semantic labeling of sub-decimeter resolution images with convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 2016.

[53] J. Wang, J. Song, M. Chen, and Z. Yang. Road network extraction: a neural-dynamic framework based on deep learning and a finite state machine. *International Journal of Remote Sensing*, 36(12):3144–3169, 2015.

[54] 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. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[55] L. Xu, T. Jun, Y. Xiang, C. JianJie, and G. LiQian. The rapid method for road extraction from high-resolution satellite images based on usm algorithm. In *2012 International Conference on Image Analysis and Signal Processing*, pages 1–6, Nov 2012.

[56] Y. Yang and S. Newsam. Bag-of-visual-words and spatial extensions for land-use classification. *ACM SIGSPATIAL*, 2010.

[57] N. Yokoya, P. Ghamisi, J. Xia, S. Sukhanov, R. Heremans, C. Debes, B. Bechtel, B. Le Saux, G. Moser, and D. Tuia. Open data for global multimodal land use classification: Outcome of the 2017 IEEE GRSS Data Fusion Contest. *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, 11(5):1363–1377, 2018.

[58] D. Zeng, T. Zhang, R. Fang, W. Shen, and Q. Tian. Neighborhood geometry based feature matching for geostationary satellite remote sensing image. *Neurocomputing*, 236:65 – 72, 2017.

[59] J. Zhao and S. You. Road network extraction from airborne lidar data using scene context. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 9–16, June 2012.

[60] X. Zhu, D. Tuia, L. Mou, G. Xia, L. Zhang, F. Xu, and F. Fraundorfer. Deep learning in remote sensing: a comprehensive review and list of resources. *IEEE Geosci. Remote Sens. Mag.*, 5(4):8–36, 2017.