Building High Resolution Maps for Humanitarian Aid and Development with Weakly- and Semi-Supervised Learning

Derrick Bonafilia
Facebook AI
dbonafilia@fb.com

David Yang
Facebook
dzyang@fb.com

James Gill
Facebook
jagill@fb.com

Saikat Basu
Facebook
saikatbasu@fb.com

Abstract

Detailed maps help governments and NGOs plan infrastructure development and mobilize relief around the world. Mapping is an open-ended task with a seemingly endless number of potentially useful features to be mapped. In this work, we focus on mapping buildings and roads. We do so with techniques that could easily extend to other features such as land use and land classification. We discuss real-world use cases of our maps by NGOs and humanitarian organizations around the world—from sustainable infrastructure planning to disaster relief. We investigate the pitfalls of existing datasets for building detection and road segmentation and highlight the way that models trained on these datasets—which tend to be highly specific to particular regions—produce worse results in regions of the world not adequately represented in the training set. We explain how we used data from OpenStreetMap (OSM) to train more generalizable models. These models outperform those trained on existing datasets, even in regions in which those models are overfit, and produce these same high-quality results for a diverse range of geographic areas. We utilize a combination of weakly-supervised and semi-supervised learning techniques that allow us to train on the noisy, crowdsourced data in OSM for building detection, which we formulate as a binary classification problem. We then show how weakly supervised learning techniques in conjunction with simple heuristics allowed us to train a semantic segmentation model for road extraction on noisy and never pixel-perfect training data from OSM.

1. Introduction

A country's census shows how many people live in a particular census tract, but it doesn't indicate where people live within a tract, and sometimes these tracts can encompass thousands of square miles. For example, Africa's largest census tract is 150,000 square miles and encompasses some 55,000 people. While highly accurate, census data is not granular enough for efforts like vaccination campaigns. These campaigns generally have limited resources that need to be allocated efficiently. Precise population distribution data facilitates and expedites humanitarian work like this by enabling organizations to locate those in need and reach them efficiently. The High Resolution Settlement Layer (HRSL), introduced by Tiecke et al. [19], combines census data with a building detection algorithm run on high resolution satellite imagery to create population density maps with 30 by 30 meter granularity. This level of granularity greatly facilitates development and relief efforts like vaccination campaigns by allowing humanitarian actors to allocate their resources to the precise areas where people live.

The building detection work discussed here builds on the work of Tiecke et al. [19] by increasing the accuracy and the geographic robustness. Using a combination of weakly-supervised and semi-supervised training techniques in conjunction with the freely available data in OpenStreetMap(OSM), we are able to locate buildings in high resolution satellite imagery [15]. Following the methodology outlined by Tiecke et al. [19], these results are joined with census data to produce extremely accurate, high resolution population density maps. Outside of the computer vision classification of patches of imagery as either building or non-building, our approach to generating population density maps is the same as in the original HRSL. Our final result is in the same format, but with higher accuracy and greater geographic coverage. The datasets resulting from this work will be released as an update to the HRSL. The release will be done region by region as we consult with interdisciplinary experts to ensure that the potential for misuse and abuse of this data is minimized and the accuracies of the resulting datasets meet the standards for release.

These maps are already having real world impact. For example, the population density map that we produced for Malawi enabled the Red Cross to quickly and remotely map around 1 million houses and 120,000 km of roads for a measles and rubella immunization campaign. This facilitated over 100,000 house visits in 3 days with just 3,000 vol-

unteers on the ground and allowed volunteers to reach many households that would have otherwise been overlooked. Our population density maps have been used for similar immunization campaigns in Mozambique by the Bill and Melinda Gates Foundation: verifying the number of children under five years old that needed polio vaccines, determining how much vaccine to procure, and validating existing vaccination coverage estimates. Additional uses of our maps include The World Food Programme's analyses on disaster preparedness and disaster response and Humanitarian OpenStreetMap Team's targeted rural electrification studies.

The other focus of our work is road segmentation. Road vectors are often at the core of user experience in mapping applications today. They make up a great deal of the visual stimulus, are the organizational system onto which addresses—and therefore homes and businesses—are mapped, and the road network they compose is used for navigation. Road segmentation does not directly lead to road vectors, but there are well established post-processing steps to extract road vectors from segmentation masks [12] [4]. In this work, we focus on the segmentation aspect of the problem. Like population density maps, road network maps allow for increased efficiency of both time and resources when responding to disasters and planning for development. Most map providers lack detailed road network information for many parts of the world, particularly in many areas of the developing world. Many of these areas are vulnerable to disasters, and when disaster strikes, the quickest way to get the detailed road information needed to provide aid and relief is through open source manual mapping, which often takes place through mapathons organized by groups like the Red Cross.

To best support humanitarian efforts, road network maps should be as accurate as possible in as many regions as possible. However, most available datasets for road segmentation are heavily biased towards particular regions. For example, the SpaceNet Roads Dataset exclusively contains data from four major cities and the DeepGlobe Roads dataset only contains data from Thailand, India, and Indonesia [1] [5]. The models that are trained on these datasets perform well in the regions well represented in the training set, and perform much worse in other regions. We use data from OSM to train more accurate and geographically robust road segmentation models. In particular, we use a threshold on the number of roads mapped in a particular area to find areas that are more completely mapped; we then use this data to train a weakly supervised road segmentation model with much greater geographic robustness. Our road segmentation models are already having real world impact. In 2018, a large area of Kerala was flooded and neither OSM nor proprietary map providers had mapped the road network sufficiently for humanitarian actors to efficiently deliver aid to those in need. A large scale human mapping effort was undertaken to provide road network data to humanitarian actors through OSM. In this instance, we released road vectors generated by our methods to the OSM community to expedite mapping of the region and therefore expedite relief to the region. We have also used these road vectors for large scale mapping efforts in Thailand and Indonesia and uploaded the results to OSM. We currently share road vectors generated through our pipeline with trusted humanitarian partners, and we are working on finding an effective and responsible way to share these road vectors with the greater humanitarian and mapping communities. While our road vectors are accurate enough to be directly used for some humanitarian efforts, we mostly use our road predictions to assist human mapping. This process allows human mappers to spend their time mapping the nuanced details of an area while our machine learning algorithm automatically maps out the road network. This leads to increased mapping efficiency, which is particularly valuable in times of crisis.

We are sharing our work with the immediate purpose of helping humanitarian actors pursue progress on the United Nations Sustainable Development Goals. We see immediate and obvious applications of this work in fighting hunger, poverty, and disease by locating where exactly populations are and which roads to take to reach these people. Work using this data has already begun with many partner organizations and we hope to continue to expand the positive impact of this work and better understand its utilities as well as its failings. We also hope that sharing this work will help highlight some of the failings of more regionally specific datasets and approaches to these problems. Using approaches that fail to generalize can negatively affect groups that fall outside of the training distribution by providing incorrect information about these areas and populations. We show empirically that there are large benefits to seeking out diverse sources of data. We hope to encourage researchers to prioritize the use of diverse sources of training data when trying to make progress on global problems.

2. Related work

Extracting information from aerial imagery has been an important research area since imagery became widely available. Due to the size of Earth, manually extracting all instances of a particular class or classifying all areas into categories is not feasible. The fields of remote sensing and Earth observation have been approaching these problems computationally for decades. Recently, deep learning has been applied to the problem. Roads and buildings are just a few of the features that deep learning has been used to compute [12] [19] . Others include crosswalks and oil palm trees [9] [2].

Data is a persistent area of attention in this research. In lieu of manually labeling training data for road detection,

Mnih and Hinton [14] investigated learning directly from data in OSM, a noisy, crowdsourced dataset. Their work developed a robust loss function to allow training on the noisy data. In contrast, we propose a processing of OSM data that allows typical Deep Neural Nets(DNNs) to be applied with typical loss functions and optimizers despite the noisiness of the data. We also approach our work at much larger scale, focusing on the geographic diversity of the training data.

Kaiser et al. [8] also investigate OSM data for training building detection and road segmentation models. They find that using OSM to pretrain their models improves segmentation results, but that using only OSM data "achieves reasonable (albeit far from optimal) results." They approach the building and road problems from the same semantic segmentation framework whereas we instead frame building detection as a classification problem. Their work is focused on urban environments where open data tends to be relatively complete. In contrast, we extend the use of OSM data beyond these urban environments to all areas of the world where buildings and roads exist and have been mapped. Maggiori et al. [10] also train convolution neural nets to detect buildings. They found that using OSM data improved results for their task, but their investigation was limited to just a few select locations.

The Functional Map of the World(fMoW) dataset also built a large dataset spanning six continents [3]. FMoW encompasses a broad range of features, whereas we focused on just two categories. FMoW has over a million annotated images and over a million annotated points of interest; its most common category is recreational facility, which has between 80,000 and 90,000 occurrences [3]. In contrast, our building detection dataset has over 50 million annotated buildings, and our road dataset has around 1.8 million images, each covering around one square kilometer and containing at least 25 roads.

3. Crowdsourced Data from OpenStreetMap

A major roadblock to scaling our maps, of both population densities and of roads, from regional projects to fully global projects has been a dearth of labeled data. As this work is unprecedented in terms of its scale, and as we found that models trained on particular regions tend to underperform in dissimilar regions, a traditional supervised learning approach to training more global models would have likely entailed manually labelling many millions more images from all over the globe—a task likely to require a large team of human annotators over a long period of time, entailing a large organizational and temporal cost. This observation, as well as the observation that OSM serves as a crowdsourced geospatial dataset of the world, led us to consider using OSM to train global models.

OSM has many labeled features, is freely available, and has data for almost all regions in the world. OSM contributors include individuals, NGOs, corporations, and many other groups. These groups have different motivations for mapping and therefore map a range of features in varying locations. This creates a diverse dataset. The regional diversity of OSM largely allows us to avoid the developed-world bias found in many other training sets; however, even though OSM has more data from the developing world than other data sources, the developed world is still overrepresented in OSM data. Additionally, using OSM data for labels has several major challenges. The first challenge is the quality and correctness of available data. Another challenge is correspondence between our imagery and OSM data. We need to ensure the OSM tags correspond temporally and spatially to our data. Another is, that in our experience, OSM's tagged features are high precision, but extremely low recall. We solve these problems in different ways for building detection and road segmentation.

4. Building Detection at Global Scale

4.1. Dataset Creation

Starting with a seed dataset of around 1M labeled patches of imagery, D, we use a combination of weakly-supervised and semi-supervised learning techniques along with data in OSM to generate a dataset, D', of more than 100 million labeled training images with an exact $50\% - 50\%$ positive and negative example split. Starting with around 1M images, we increase the total ground area labeled in our training set from around 1 billion square meters of imagery to around 100 billion square meters distributed as 1000 square meter patches across 79 distinct regions spanning 6 continents and a massive range of cultural and architectural styles. Though this process is automated, it does rely on a small amount of manually labeled data for a given region in order to create the larger training set for that region. In our case, we create a dataset with 100 images for every 1 image manually labeled. Figure 1 gives a visualization of how this process works for Great Britain, a relatively densely populated region where both the weakly-supervised and semi-supervised data collection steps play important roles. More details on D and D' are in the Supplementary Materials.

Weakly Supervised Approach We approach the problems of data correctness and spatial and temporal alignment through our weakly supervised approach to collecting positive examples. We solve these problems simultaneously by labeling an image as a positive example if, and only if, there is a house in that image according to a label in OpenStreetMap and our pre-processing step detected straight edges in that image. The pre-processing step is the same as described in Zhang et al. [21]. In our experi-

ments, the images found through this approach in a region of the Sahara where buildings are quite sparse and where we would expect this method to have the most difficulty, were correctly labeled as having houses for 996 out of 1000 images[99.6% accuracy]. According to the analysis done by Zhang et al. [21], this is more accurate than the human labelers they used for data collection. Despite the high accuracy of our labels, we acknowledge the possibility of systematic errors and biases in the labels accrued through OpenStreetMap due to issues such as imagery misalignment or incorrect mapping. Thus, we characterize our approach as weakly rather than fully supervised learning.

Semi-Supervised Approach The low recall of OSM feature tags makes the collection of negative examples more complex. The lack of a labeled building could mean there is no building there, but it could also mean that the area has not been mapped yet, in which case it gives no insight into whether or not we should label this image as containing a building. Treating all areas without tagged buildings as non-building images provides only a slight bias towards returning non-building images. Here, we rely on a semi-supervised bootstrapping approach along with a simple statistical approach to bound the expected error rate of our non-building labels to below 1 percent. We note that this bootstrapping combined with processing has similarities with the data distillation work done by Radosavovic et al. [17].

We use the outputs of our 18 layer Resnet that we've trained on D, our original one million manually labeled images, to find examples of tiles not containing buildings for our new training set, D' [6]. For each region, we set a threshold, τ, for the output— above this threshold, images are considered to contain a building, and below this threshold, they are not. For our negative samples, we uniformly sample from the images that we predicted to be negative. Given the outputs of our 18 layer Resnet, p, on a validation set with ground truth labels, l—the number of false negatives(FN), the number of predicted negatives(PN), and the f_1 score are all a function of p, τ, and l. Since p and l are fixed for a given region, FN, PN, and the f_1 score are all functions of τ. Our expected labeling error in a given region is therefore FN / PN. To ensure an expected labeling error less than 1 percent, we set τ to maximize the f_1 score subject to the constraint that our $FN/PN < .01$. That is:

Given a fixed p and l:

$$f(\tau) = FN, g(\tau) = PN, h(\tau) = f_1 \tag{1}$$

$$\tau = \underset{x \in [0,1]}{\arg\max}\, h(x) \mid f(x)/g(x) < .01 \tag{2}$$

Then, for each region, for each image with no houses tagged in OSM, we use the outputs from our 18-layer Resnet to score it as building or non-building using τ as our threshold. If we have pulled x labeled houses from a given region, we then randomly sample x non houses from the region to create our new training set, D', with a 50-50 building/non-building split.

Figure 1. Satellite images from Great Britain. From left to right, images where we've detected edges(but there are no mapped buildings in OSM), images our approach labels as containing buildings, images our approach labels as not containing buildings.

Figure 2. Road extraction from satellite imagery in Mexico.

4.2. Training

First we trained 18, 34, and 50 layer Resnets on our seed dataset, D [6]. On these seed set, we found overfitting to negatively affect the performance of the 34 and 50 layer Resnets and our best performing model was the 18 layer Resnet. We use this model as a baseline with which to compare the models trained on our new dataset D'. Then, using around 30 million of our newly collected dataset, D', from our weakly-/semi-supervised approach (around 400,000 labeled images from each of 78 regions), we trained 18, 34, and 50 layer Resnets. As expected the 50-layer Resnet was the best performing model of this group. We then explore using the initial seed dataset, D, to finetune this model and see even larger accuracy improvements.

4.3. Results

We compare our approach to two baselines. The first baseline is the fully trained model produced by Zhang et al. [21]. This model is intended to be a fully global building detection model, but was trained on a more limited dataset using a weakly supervised semantic segmentation-based approach to classification. It is the state of the art in the literature for building detection from satellite imagery and utilizes the same imagery source and testing methodology. It therefore seems the best comparison of our approach to existing approaches in the literature. The second baseline is

Table 1. Number of regions in which approach on the left outperforms approach on top. "Finetune" signifying first training on D' and then finetuning on D.

	Zhang et al. [21]	D	D'	Finetune
Zhang et al. [21]		6	6	5
D	67		14	4
D'	67	59		18
Finetune	68	69	55	

Table 2. Number of regions with above a certain f_1 score for each approach

	.2	.4	.6	.8	.85	.9	.95
Zhang et al. [21]	73	71	68	55	41	11	2
D	73	73	73	72	67	52	5
D'	73	73	73	73	69	53	6
Finetune	73	73	73	72	70	57	10

the 18-layer Resnet that we trained on our seed dataset D.

There are 73 distinct geographic regions for which we have test data for both Zhang et al. [21]'s work and our own work. We first compare our two baselines with our model trained on our new OSM-based dataset D'. To focus on both generalizability and accuracy we focus on two accuracy measurements: in how many regions each approach provided the highest f_1 score and the mean f_1 score across regions, weighting each region equally. We use the f_1 score since this is a strongly class imbalanced classification problem. The model trained on D' gets the best f_1 score on 53 regions, the model trained on D gets the best f_1 score on 14 regions, and Zhang et al. [21]'s model performs the best on 9 regions. The largest improvement in accuracy is from Zhang et al. [21]'s work to our baseline with our seed dataset with an increase of mean f_1 from .818 to .907. However, there is still a large accuracy improvement from our baseline to our OSM-trained model with the mean f_1 score improving further to .914. We then finetune our model trained on our OSM-based dataset D' with the manually labeled data in D and see a further improvement of mean f_1 to .920. More detailed results are shown in tables 1 and 2 and in the Supplementary Material.

5. Road Segmentation at Global Scale

Mahajan et al. [11] show that, for classification tasks, deep neural nets trained on large datasets are robust to, yet still negatively effected by, label noise. In this work, we show the same robustness for the task of semantic segmentation. We first process OSM data into a training dataset using simple heuristics. We then train a modified version of the DLinkNet-34 model that won 2018's DeepGlobe challenge on this data[22] [5].

5.1. Dataset Creation

The structure of the building classification problem as well as the test sets we had available for each region allowed for a conceptually simple approach to creating a training set with a bounded expected labeling error rate. The structure of semantic segmentation, due to pixel level rather than image level labels is not obviously amenable to that sort of approach. Thus, rather than try to construct a training dataset with certain guaranteed desirable properties, we started with a naive approach.

We structure our dataset to minimize the complexity of composing the dataset while also attempting to retrieve relatively high quality training labels. OSM represents roads as vectors. Most road segmentation work attempts to get per pixel raster labels that exactly match road contours; this type of label is labor intensive to generate. To keep the dataset generation simple, we rasterize each edge of each road vector to 5-pixel-width lines, noting that this label is almost never fully correct. We show in figure 2 that even though we use the same pixel width for all training labels, the model learns to predict roads that match the more complex contours and varying widths of the roads shown in the imagery. We handle the removal of low quality training areas due to incomplete map data in similarly naive fashion. We first tile the world using the Bing Map Tile system [16]. We collect our training set at zoom level 15, which we represent as 2048 by 2048 pixel input images. We throw out all tiles where less than 25 roads have been mapped, as we found tiles with fewer roads often only mapped out major roads . For each remaining tile we rasterize the road vectors as explained above and use the resulting mask as our training label. To work at the same resolution as the DeepGlobe dataset, we use a random 1024 by 1024 pixel crop from the initial 2048 x 2048 data and labels. At the time of our OSM snapshot, we find around 1.8 million tiles that meet our heuristic of having at least 25 roads mapped. This adds up to more than 1.8 million square kilometers of land area coverage, an increase of more than 1000X from the 1,632 square kilometers of data in the DeepGlobe dataset. The geographic distribution of this dataset is shown in the Supplementary Material.

5.2. Training

We train using the DLinkNet34 model that won last year's DeepGlobe challenge. We train using SGD rather than Adam as we found it generalized better. We initially found that training the model on just one region at a time using the OSM-based dataset worked, but that attempting to train on several regions or all available regions failed to converge. As noted by Wu and He [20], Batch Normalization does not perform well on small batches and our large input image size required small batches [7]. After switching Batch Normalization layers to Group Normaliza-

tion layers, we found that the more regions in the training set, the better the model performed, strengthening our belief that more geographic diversity improves model performance. We trained on satellite imagery from almost all regions in the world that had some areas of relatively complete OSM data. After training the model on the OSM data, we also investigate finetuning the model on the DeepGlobe dataset and produce results that are state of the art on the road extraction challenge [5].

5.3. Results

We first evaluate our model on the DeepGlobe road extraction challenge. Having only trained on weakly supervised OSM data, our model receives a road IOU score 0.565 on the validation set. This is an improvement over the baseline put forward in the challenge [5]. Unlike most entrants to the challenge, we run our model only once per image with no Test Time Augmentation(TTA) or post-processing; despite this, we find that our model is competitive with other entries. It is outperformed by the best entrants, like the original DLinkNet34 submission, but outperforms many other entrants[5]. Our model is the first trained solely on OSM data to be competitive with road segmentation models trained in a fully supervised manner. This result is more meaningful because the models trained on DeepGlobe are being evaluated on the very region to which the DeepGlobe dataset is overfit. We then test our finetuned model on the same dataset, again with a single pass and no TTA or post-processing and produce a mean road IOU score of 0.6352. The winning submission, the DLinkNet34 model that we have adopted, receives a road IOU score of 0.6466, but attributes 0.029 of that IOU to TTA alone. This brings their single pass, no TTA score down to 0.6176. Our model thus outperforms the winning submission of the DeepGlobe challenge in single pass accuracy by around 2%. This falls in line with other literature showing that pretraining on OSM data and finetuning on a manually labeled dataset improves accuracy [8] [10]. As we discuss later, the globally trained model outperforms regionally trained models by a large margin on regions outside of the regional training distribution of the regionally trained models.

6. Comparison to existing Datasets

In the building detection problem, we reconfirmed the findings of Zhang et al. [21] that, even with a large dataset, the more regions involved in the training set, the better the model does on all regions. For example, the model trained on data from all 78 regions outperforms a model trained on millions of labels from only 9 Sub-Saharan African regions on every single region, including those 9. In road segmentation, we also find that diversity of data appears to be more important than the quantity of data. Models trained on the global OSM data outperform models trained on OSM data from Thailand, Indonesia, and India alone.

Table 3. IOU score of Zhou et al. [22]'s DeepGlobe-trained DLinkNet34 and our OSM-trained DLinkNet34 on various datasets outside the DeepGlobe datasets geographic training distribution.

Dataset	DG IOU	OSM IOU
Spacenet - Paris	.161	.324
Spacenet - Vegas	.172	.425
Spacenet - Shanghai	.171	.249
Spacenet - Khartoum	.184	.312
Spacenet - Mean	.172	.328
Mnih [13]	.399	.464
Mean of all	.218	.355

For road segmentation, we qualitatively show the difference between a globally trained model and a model trained on the DeepGlobe regional data in the Supplementary Material. We quantitatively show the difference in generalizability of the two approaches by evaluating each of the models on the road detection test set proposed in Mnih [13] and on each of the regions of the Spacenet Roads dataset with the labels rasterized as in Singh et al. [18] [1]. The model trained on global OSM data outperforms the model trained on DeepGlobe data on all regions. Across these 5 test sets, the mean IOU score of the DeepGlobe model is .218 and the mean IOU score of the OSM trained model is .355 with a 62 percent relative improvement and a 13.7 percent absolute improvement. The results are shown in full in table 3.

There is a clear trend of increased diversity of training data increasing both accuracy and generalizability, yet all benchmark datasets are highly specific to certain regions. Developing more diverse datasets is crucial to increasing performance and understanding how models perform in different parts of the world.

7. Conclusion

Our work is the first to consider using OSM pretraining on a global rather than regional scale. Furthermore, our work is the first to show that training on OSM alone (and not finetuning at all) provides high enough quality results to use in development and relief efforts. This creates a clear path forward expanding road segmentation and building detection models from regions in which they work well into regions in which they do not. One option is training exclusively on global OSM data. A path to even stronger results is pretraining on global OSM data and finetuning on a small amount of manually labeled data for regions of interest.

References

[1] Spacenet on amazon web services (aws). *Datasets. The SpaceNet Catalog.*, Last modified April 30, 2018. URL https://spacenetchallenge.github.io/datasets/datasetHomePage.html. 2, 6

[2] R. F. Berriel, A. T. Lopes, A. F. De Souza, and T. Oliveira-Santos. Deep learning-based large-scale automatic satellite crosswalk classification. *IEEE Geoscience and Remote Sensing Letters*, 14(9):1513–1517, 2017. 2

[3] G. Christie, N. Fendley, J. Wilson, and R. Mukherjee. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6172–6180, 2018. 3

[4] İ. Demir, F. Hughes, A. Raj, K. Dhruv, S. Muddala, S. Garg, B. Doo, and R. Raskar. Generative street addresses from satellite imagery. *ISPRS International Journal of Geo-Information*, 7(3):84, 2018. 2

[5] I. Demir, K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia, and R. Raska. Deepglobe 2018: A challenge to parse the earth through satellite images. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 172–17209. IEEE, 2018. 2, 5, 6

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[7] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 5

[8] P. Kaiser, J. D. Wegner, A. Lucchi, M. Jaggi, T. Hofmann, and K. Schindler. Learning aerial image segmentation from online maps. *IEEE Transactions on Geoscience and Remote Sensing*, 55(11):6054–6068, 2017. 3, 6

[9] W. Li, H. Fu, L. Yu, and A. Cracknell. Deep learning based oil palm tree detection and counting for high-resolution remote sensing images. *Remote Sensing*, 9(1):22, 2016. 2

[10] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez. Convolutional neural networks for large-scale remote-sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(2):645–657, 2017. 3, 6

[11] D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, and L. van der Maaten. Exploring the limits of weakly supervised pretraining. *arXiv preprint arXiv:1805.00932*, 2018. 5

[12] G. Máttyus, W. Luo, and R. Urtasun. Deeproadmapper: Extracting road topology from aerial images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3438–3446, 2017. 2

[13] V. Mnih. *Machine learning for aerial image labeling*. University of Toronto (Canada), 2013. 6

[14] 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. 3

[15] OpenStreetMap contributors. Planet dump retrieved from https://planet.osm.org . https://www.openstreetmap.org, 2018. 1

[16] X. Qu, M. Sun, C. Xu, J. Li, K. Liu, J. Xia, Q. Huang, C. Yang, M. Bambacus, Y. Xu, et al. A spatial web service client based on microsoft bing maps. In *2011 19th International Conference on Geoinformatics*, pages 1–5. IEEE, 2011. 5

[17] I. Radosavovic, P. Dollár, R. Girshick, G. Gkioxari, and K. He. Data distillation: Towards omni-supervised learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4119–4128, 2018. 4

[18] S. Singh, A. Batra, G. Pang, L. Torresani, S. Basu, M. Paluri, and C. Jawahar. Self-supervised feature learning for semantic segmentation of overhead imagery. 6

[19] T. G. Tiecke, X. Liu, A. Zhang, A. Gros, N. Li, G. Yetman, T. Kilic, S. Murray, B. Blankespoor, E. B. Prydz, and H. H. Dang. Mapping the world population one building at a time. *CoRR*, abs/1712.05839, 2017. URL http://arxiv.org/abs/1712.05839. 1, 2

[20] Y. Wu and K. He. Group normalization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018. 5

[21] A. Zhang, X. Liu, A. Gros, and T. Tiecke. Building detection from satellite images on a global scale. *CoRR*, abs/1707.08952, 2017. URL http://arxiv.org/abs/1707.08952. 3, 4, 5, 6, 9

[22] L. Zhou, C. Zhang, and M. Wu. D-linknet: Linknet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 182–186, 2018. 5, 6

Supplementary Materials

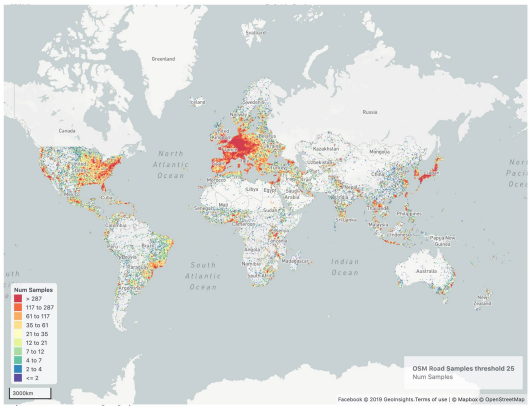

Figure 3. Visualization of the geographic distribution of training data for the OSM road segmentation model. Missing areas are due to satellite imagery being unavailable at the time of experiments.

Figure 4. Road extraction from satellite imagery in rural Mexico. Left: Satellite Imagery. Middle: THA/IND/IDN trained model. Right: Global OSM trained model. The model trained on DeepGlobe data misses the road in the top left almost entirely and leaves several roads in the middle of dense trees whereas the globally trained model performs well.

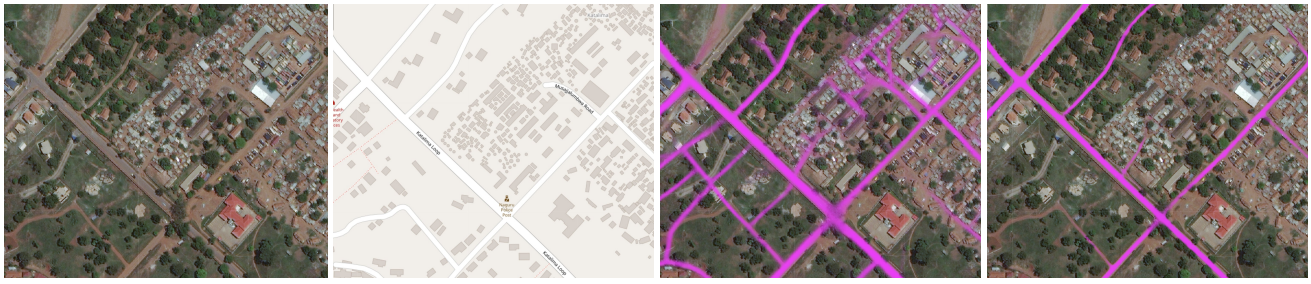

Figure 5. Road extraction from a relatively well mapped area in Kampala, Uganda. From left to right: satellite imagery, OSM(manually mapped), THA/IND/IDN trained model, Global OSM trained model. The model trained on DeepGlobe draws numerous non-existent roads through the middle of houses whereas the globally trained model performs well.

Table 4. Number of training samples in given regions for datasets D and D'. Information about some regions withheld until they go through a data release clearance process.

ISO3	Samples in D	Samples in D'
AGO	9999	336978
ARG	0	400000
AUT	0	400000
BEL	21281	400000
BEN	8347	400000
BFA	12687	0
BGD	9429	400000
BWA	235	400000
CAF	19999	400000
CHL	38857	189079
CIV	10059	0
CMR	38987	0
COL	0	400000
DOM	0	135776
DZA	28611	0
EGY	21342	0
ESP	0	400000
EST	19399	400000
FIN	20000	400000
FRA	19602	400000
GBR	19478	400000
GHA	12462	0
GIN	0	400000
GRC	0	400000
GTM	0	400000
HKG	0	64629
HND	0	321302
HTI	25377	0
IDN	3283	400000
IRL	10000	400000
ISL	39492	112124
ITA	0	400000
JPN	0	400000
KAZ	0	400000
KEN	7502	0
KHM	17926	346119
KOR	0	384946
LBR	0	400000
LKA	2110	0
LSO	19418	400000
LTU	18504	400000
LUX	0	264771
LVA	1990	400000
MDG	43241	9657
MEX	19491	0
MLI	0	400000
MOZ	21379	285596
MRT	0	117375
MWI	23789	0
MYS	0	400000
NER	29342	377679
NGA	21829	400000
NOR	39944	400000
NPL	14713	400000
NZL	0	400000
PER	30515	11461
PHL	5498	0
PNG	19702	75867
POL	0	400000
PRI	0	400000
RWA	25384	400000
SEN	29235	0
SLE	0	400000
SLV	0	79790
SWE	19777	400000
TCD	20501	400000
THA	24018	0
TUN	0	114604
TWN	18208	157117
TZA	18689	400000
UGA	13661	400000
UZB	37764	0
VNM	46143	361541
ZAF	17339	400000
ZMB	10000	400000

Table 5. f_1 score of the various approaches on the 73 regions for which we had test sets for all approaches. Seed Dataset D' refers to the Resnet18 trained on D', OSM Dataset D' refers to the Resnet50 trained on D', and the final column refers to the Resnet50 first trained on D and then finetuned on D'. One region name withheld as it has not gone through the process to be released publicly.

ISO3	Zhang et al. [21]	Seed Dataset D'	OSM Dataset D'	Pretrain D and Finetune D'
AGO	0.812	0.905	0.906	0.928
ARG	0.898	0.951	0.956	0.958
AUT	0.846	0.929	0.940	0.942
BEL	0.732	0.855	0.853	0.857
BEN	0.887	0.926	0.936	0.931
BGD	0.855	0.872	0.877	0.902
BWA	0.788	0.887	0.899	0.909
CAF	0.761	0.926	0.929	0.937
CHL	0.873	0.936	0.966	0.951
COL	0.872	0.926	0.919	0.930
DOM	0.915	0.891	0.889	0.903
ESP	0.859	0.929	0.943	0.942
EST	0.631	0.836	0.835	0.846
FIN	0.555	0.831	0.852	0.854
FRA	0.765	0.874	0.884	0.889
GBR	0.675	0.844	0.843	0.852
GIN	0.842	0.918	0.936	0.934
GRC	0.883	0.929	0.946	0.948
GTM	0.885	0.935	0.938	0.939
HKG	0.866	0.897	0.901	0.909
HND	0.889	0.900	0.907	0.915
IDN	0.933	0.911	0.914	0.918
IRL	0.672	0.862	0.860	0.875
ISL	0.333	0.780	0.784	0.800
ITA	0.856	0.926	0.935	0.938
JPN	0.870	0.924	0.925	0.928
KAZ	0.901	0.908	0.892	0.914
KOR	0.815	0.835	0.844	0.856
LBR	0.928	0.907	0.903	0.920
LSO	0.767	0.869	0.883	0.868
LTU	0.801	0.910	0.908	0.913
LUX	0.863	0.926	0.928	0.936
LVA	0.724	0.893	0.897	0.908
MLI	0.852	0.940	0.946	0.953
MRT	0.829	0.915	0.912	0.924
MYS	0.897	0.906	0.914	0.925
N/A	0.880	0.942	0.941	0.945
NER	0.782	0.912	0.920	0.929
NGA	0.853	0.931	0.936	0.942
NOR	0.287	0.874	0.859	0.879
NPL	0.816	0.924	0.927	0.933
NZL	0.858	0.940	0.938	0.947
PER	0.922	0.941	0.949	0.947
PNG	0.854	0.917	0.898	0.908
POL	0.843	0.932	0.924	0.928
PRI	0.919	0.951	0.960	0.964
SEN	0.879	0.955	0.970	0.976
SLE	0.896	0.913	0.928	0.924
SLV	0.887	0.909	0.914	0.917
SWE	0.571	0.877	0.849	0.863
TCD	0.611	0.825	0.837	0.845
TUN	0.906	0.925	0.937	0.943
TWN	0.821	0.871	0.890	0.892
UGA	0.829	0.923	0.916	0.924
VNM	0.948	0.938	0.947	0.950
ZMB	0.715	0.903	0.901	0.914