Crowd-Sourced Road Quality Mapping in the Developing World

Benjamin Choi[1] , **John Kamalu**[1]

[1]Stanford University

{benchoi, jkamalu}@stanford.edu

Abstract

Road networks are among the most essential components of a country's infrastructure. By facilitating the movement and exchange of goods, people, and ideas, they support economic and cultural activity both within and across borders. Up-to-date mapping of the the geographical distribution of roads and their quality is essential in high-impact applications ranging from land use planning to wilderness conservation. Mapping presents a particularly pressing challenge in developing countries, where documentation is poor and disproportionate amounts of road construction are expected to occur in the coming decades. We present a new crowd-sourced approach capable of assessing road quality and identify key challenges and opportunities in the transferability of deep learning based methods across domains.

1 Introduction

As the world rapidly urbanizes, road networks follow suit, enabling the transport of essential services and resources to support this development. This growth is expected to occur disproportionately in the developing world, with non-OECD countries expected to account for 90% of global road infrastructure development over the next 40 years and African "development corridors" expected to add upwards of 53,000 km of new roads [Dulac, ; Laurance *et al.*, 2015]. Roads associated with rapid development are often poorly documented or illegal–there are roughly 3 km of illegal or unofficial roads for every 1 km of legal roads in the Brazilian Amazon [Alamgir *et al.*, 2017].

Road quality is a commonly used proxy for tracking economic activity and development [Fan and Chan-Kang, 2005; Gertler *et al.*, 2016]. Policymakers and conservationists therefore often rely on up-to-date maps of road quality to inform where resources can most effectively be applied. Recent advances in deep learning have adapted models originally developed for biomedical image segmentation to achieve strong performance in automatic road extraction (i.e. segmentation) tasks from satellite imagery [Zhang *et al.*, 2017]. However, the binarized outputs of these models tend to preserve only structural features, otherwise treating a bustling highway and

rural road equivalently. Therefore, road quality classification has been layered on top of segmentation outputs as a more explicit development indicator. Recent road classification methods [Cadamuro *et al.*, 2019; Cadamuro *et al.*, 2018] have relied on data measured according to the International Roughness Index (IRI), a measurement scheme that requires sophisticated instruments and technical expertise that is difficult and expensive to maintain for a considerable amount of road coverage. Different countries also feature highly varied terrain, raising the possibility of poor performance in out-of-country domains.

We extend current methods in road quality classification to leverage accessible crowdsourced data from OpenStreetMap [OpenStreetMap contributors, 2017]. We use OpenStreetMap to sample satellite imagery along a diverse set of roads of varying quality from multiple developing countries. This imagery is used to train a deep learning road quality classification model. We then apply data masks to highlight the influence of context on model performance. Such methods enable us to anticipate and address the challenges of multi-domain applications of modern classification models. We therefore propose two primary contributions: 1.) the development of a crowd-sourced approach to road quality classification and 2.) a new strategy for diagnosing issues and improving performance in out-of-country domain applications.

2 Methodology

2.1 Data

Our input data consists of 30 cm resolution satellite imagery from DigitalGlobe [Demir *et al.*, 2018] combined with road shape information and classification labels derived from OpenStreetMap. In particular, one road-label pair will contain: one $1000px^2$ satellite image covering a $300m^2$ area centered on a given road, a classification of the road belonging to the label set {major, minor, two-track}, and GPS coordinates along each road (the polyline). Each classification consists of multiple aggregated subclassifications to increase the resilience of crowd-sourced labels to varying standards of classification from country to country. Figure 1 contains representative examples of each road type. A train:test split of 9:1 is used.

We choose Kenya and Peru as our country domains based on their development status and high potential for
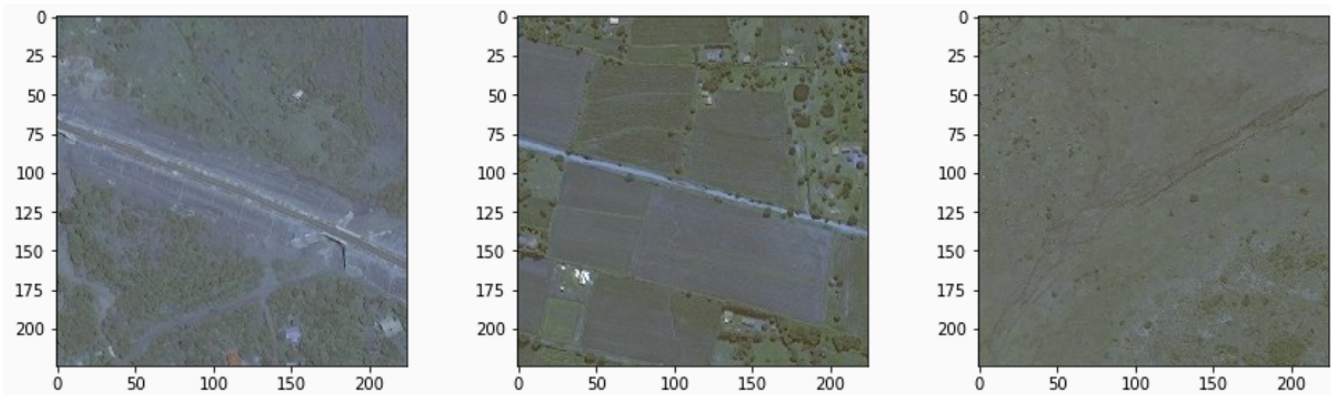
Figure 1: Example RGB images of three road quality classes: major, minor, two-track (left to right).

future road expansion. Points are sampled from coordinates supplied from OpenStreetMap along each road. We therefore generate 170K images in Kenya with a 1:20:12 `major:minor:two-track` ratio and approximately 97K images in Peru with a 1.3:2:1 ratio. We also enforce a simple threshold on RGB band values to remove images occluded by clouds.

2.2 Crowd-Sourcing Road Quality

To speed up training, we use pre-trained ImageNet [Deng *et al.*, 2009] classifiers as a foundational model for transfer learning. We append a feed-forward network to these models to translate the output embeddings of "headless" models into a final classification of road quality.

As a baseline for deep convolutional classifiers, we choose a standard ResNet model [He *et al.*, 2016]. For all models, we optimize using categorical cross-entropy loss and an Adam optimizer with a learning rate of 0.0001. After fine-tuning our baseline ResNet model on de-clouded data (using only Kenyan images, for speed of development), we try other CNN architectures and find that in practice, Xception [Chollet, 2016], a model based on depthwise separable convolutions, achieves strong performance. Therefore, all non-baseline experiments use the Xception model. These experiments also train and test on images from the same country.

Finally, we evaluate the generalizability of our model using an out-of-country domain. Therefore, in contrast to the previous experiments, we use Kenya images for training and Peru images for testing, and vice versa.

2.3 Strategies for Diagnosing Domain Transferability

A serious limitation of modern road classification methods is generalizability to disparate domains (i.e. different countries or regions), as expensive recollection of data and retraining therefore becomes necessary to accommodate the domain shift. We seek to identify and mitigate the primary challenges in out-of-country domain performance. We note that our images consist of both road surfaces and surrounding terrain. We therefore investigate the effect of each image component in isolation. In particular, we are interested in assessing the extent to which modern road classification methods rely on surrounding geographical context for classification, as this may drastically affect model performance due to highly varied geographical conditions globally.

A preliminary step in diagnosing overreliance on geographical context in road classification is a simple binary mask. Our binary mask generation process is as follows. For each Kenyan satellite image, a set of randomly geographical coordinate points within the image are provided. Since these points are randomly sampled in a larger bounding box from which the image is cropped, the number of points relevant to the road segment within the image is variable. Coordinates are connected and converted to whole pixel coordinates using the Bresenham line algorithm [Bresenham, 1965]. Each mask pixel is then expanded radially to fully cover the area of the image representing roads. Pixel-wise multiplication of our raw satellite image with our generated mask therefore results in context-occluded images. We also invert the road mask to train only-context models for comparison. The performance of our model in differing masking scenarios acts as a proxy for the extent to which our models utilize actual road pixels in classification of the overall image. We use a balanced dataset for training to eliminate unwanted effects from class imbalances.

3 Results

	Accuracy	F1
Peru - No Mask	67%	66%
Kenya - No Mask	80%	73%
Kenya - Context Occluded	77%	72%
Kenya - Road-Occluded	79%	71%

Table 1: Results from masking experiments. Accuracies reported here are unweighted.

Train	Test	Accuracy
Kenya	Peru	46%
Peru	Kenya	60%

Table 2: Results from out-of-country domain experiments.

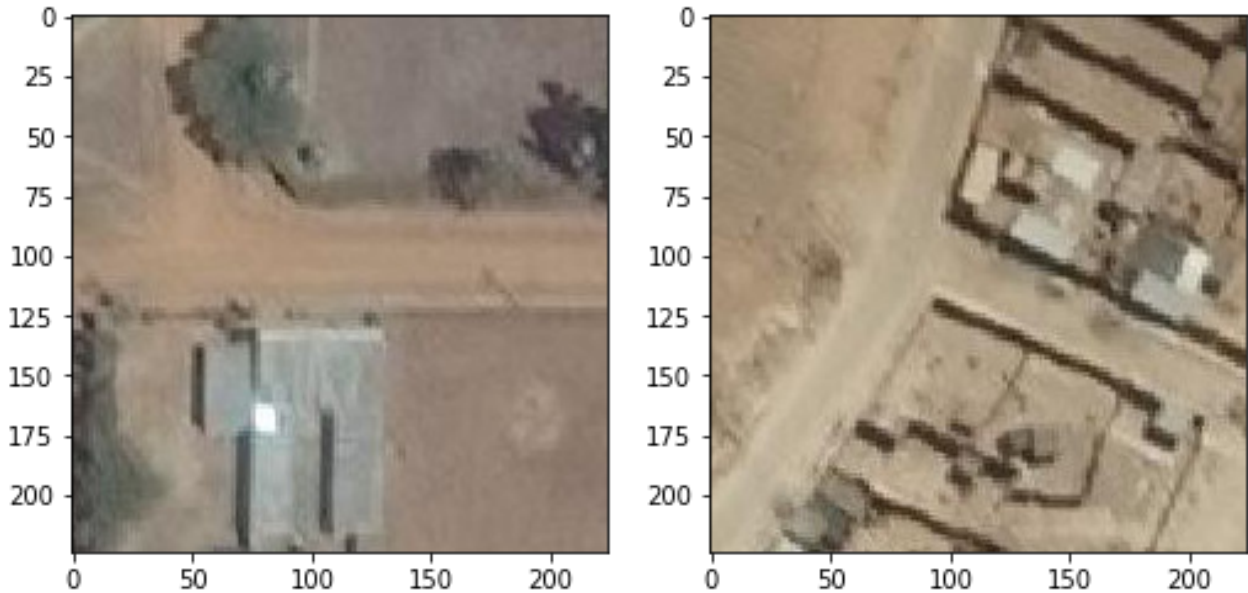**Crowd-Sourcing Road Quality:** Our baseline ResNet

Figure 2: A Kenyan minor road classified as a two-track road by Peru-to-Kenya model (left) and a true Peruvian two-track road (right). Their similarity conveys the difficulty in accommodating domain shifts between regions.

model achieves 70% classification accuracy on Kenyan images. In Table 1, we find that a pre-trained Xception model outperforms this baseline considerably in the *no mask* scenario. Relatively strong performance across road types and countries suggests that the use of crowd-sourced labels may be a vastly more scalable alternative to IRI data collection.

Based on Table 2, we find that there is a significant decrease in model performance when applying Kenya-trained models to Peru, while there is a lesser decrease when applying Peru-trained models to Kenya. The variability of performance based on direction of transfer suggests that some countries may have country-specific signals (e.g. Kenya) relevant to classification, while others may have more general signals (e.g. Peru), possibly allowing for training models more ideally-equipped for robust classification across domains. Our results indicate that implementation at scale will require relatively granular, localized training data specific to regions of interest in order to accurately capture associations with context. In contrast to traditional methods of technically-advanced IRI measurements, crowd-sourced measurements can be captured and shared by anyone at a hyperlocal resolution. The strength of crowd-sourced methods is therefore emphasized by cross-domain experiments.

While simply masking context may seem like a trivial solution to avoiding reliance on country-specific context, this is undesirable for several reasons. First, the appearance of roads in different classifications are country-specific, as overhanging vegetation, vehicle types, and even soil color (e.g. for two-track roads) all have strong associations with specific locales. Further, when road quality is used as a proxy for development, the state of the surrounding built environment may be a desired component in the overall model's assessment of the extent of urbanization. The crowd-sourced road quality

classification models presented here may act as a benchmark as the use of crowd-sourced data becomes increasingly viable and advantageous over manually collecting data.

Strategies for Diagnosing Domain Transferability: Table 1 also summarizes results from each masking approach. Both models achieve similar performance to our non-occluded baseline, with road-occluded models performing slightly better than context-occluded models. These results support the notion that road classification models tend to rely heavily on context when available, and are even able to classify "roads" with reasonable performance without the actual road being present in the image. The comparable performance of road-occluded models also suggests that road quality classifications may have strong associations with geographical context. Our findings suggest that there is a strong correlation between road type and surrounding geographic context. However, our context-occluded models still achieve reasonable performance, suggesting that the road itself does produce a signal relevant to the model.

The relative invariance of our model to occlusion of context and road can be better understood through the use of class activation maps [Zhou *et al.*, 2015]. As shown in Figure 3, class activation maps highlight per-class discriminative image regions. While major roads do seem to focus on relevant road areas of the image, minor and two-track roads often seem to rely on contextual information for classification. The occlusion invariance of our model may therefore be attributed to differences in model attention regions across classes.

The results presented here do not imply that extremely localized image-label pairs are necessary to achieve a model that can accurately associate the unique context of a region with road types. While it does suggest that considerably greater investment in crowd-sourcing up-to-date labels is nec-
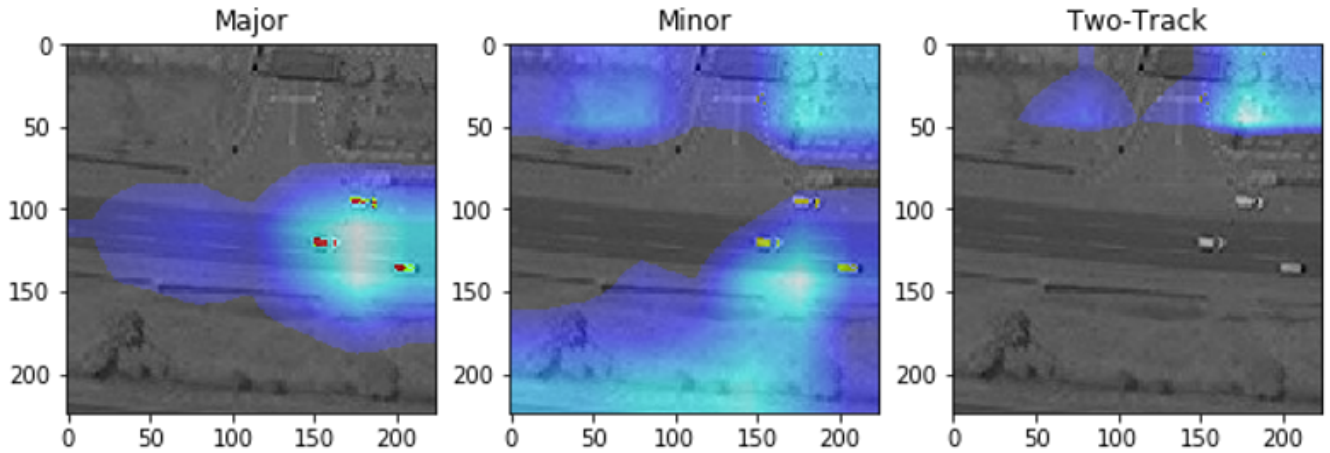
Figure 3: Class activation maps generated for a Kenyan major road reveal disparate activation regions depending on the classification.

essary, at some intermediate resolution, models should be able to generalize effectively to unseen images.

4 Conclusion and Future Work

We have shown that models that allow for the use of surrounding context in road classification may be more susceptible to failure cases in applications to out-of-country or out-of-region applications. Surrounding vegetation, land use, building styles, and other contextual factors likely possess at least a weak association with the specific country or region of interest. Careful curation of datasets used for training road and infrastructure classification models is consequently essential. Depending on the application, sampling of datasets should occur either within close local proximity of the region of interest, or with diverse multi-region input for more generalized applications. It is important to note that the adequacy of crowd-sourced labels as a proxy for in-situ road quality measurements has not been formally verified beyond visual inspection. However, in rapidly developing regions with no other road quality measurement system, our methods represent a useful first pass in tracking global road quality.

The development of methods that leverage crowd-sourcing is therefore crucial in ensuring equitable, democratized use of data that can strongly influence the course of events in some of the most rapidly developing regions of the world. We have seen that models struggle to generalize to multiple domains, necessitating the collection of localized data. In regions with such rapid development, relying on sparse, out of date measurements (e.g. IRI) is both subpar and unnecessary. As internet access becomes more widespread, the power of local infrastructure development tracking may be shifted through crowd-sourcing from governmental and scientific institutions directly to people living through this change.

Future work may address the temporal dimension of road quality tracking. While we have highlighted the challenges posed by variability across the spatial dimension in the image domain, it is also possible that anthropogenic climate change and other ecologically-influential forces may exert strong shifts in the image domain over time. By accounting for the temporal dimension, it may be possible to not only track real-time road infrastructure development, but to predict its future course.

References

[Alamgir *et al.*, 2017] Mohammed Alamgir, Mason J Campbell, Sean Sloan, Miriam Goosem, Gopalasamy Reuben Clements, Mahmoud I Mahmoud, and William F Laurance. Economic, socio-political and environmental risks of road development in the tropics. *Current Biology*, 27(20):R1130–R1140, 2017.

[Bresenham, 1965] J. E. Bresenham. Algorithm for computer control of a digital plotter. *IBM Syst. J.*, 4(1):25–30, March 1965.

[Cadamuro *et al.*, 2018] Gabriel Cadamuro, Aggrey Muhebwa, and Jay Taneja. Assigning a grade: Accurate measurement of road quality using satellite imagery. *arXiv preprint arXiv:1812.01699*, 2018.

[Cadamuro *et al.*, 2019] Gabriel Cadamuro, Aggrey Muhebwa, and Jay Taneja. Street smarts: Measuring intercity road quality using deep learning on satellite imagery. In *Proceedings of the 2Nd ACM SIGCAS Conference on Computing and Sustainable Societies*, COMPASS '19, pages 145–154, New York, NY, USA, 2019. ACM.

[Chollet, 2016] François Chollet. Xception: Deep learning with depthwise separable convolutions. *CoRR*, abs/1610.02357, 2016.

[Demir *et al.*, 2018] 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, 2018.

[Deng *et al.*, 2009] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[Dulac,] John Dulac. Global land transport infrastructure requirements.

[Fan and Chan-Kang, 2005] Shenggen Fan and Connie Chan-Kang. *Road development, economic growth, and poverty reduction in China*, volume 12. Intl Food Policy Res Inst, 2005.

[Gertler *et al.*, 2016] Paul J Gertler, Marco Gonzalez-Navarro, Tadeja Gracner, and Alexander D Rothenberg. Road quality, local economic activity, and welfare: Evidence from indonesia's highways. 2016.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. *CoRR*, abs/1603.05027, 2016.

[Laurance *et al.*, 2015] William F. Laurance, Sean Sloan, Lingfei Weng, and Jeffrey A. Sayer. Estimating the environmental costs of africa's massive "development corridors". *Current Biology*, 25(24):3202–3208, December 2015.

[OpenStreetMap contributors, 2017] OpenStreetMap contributors. Planet dump retrieved from https://planet.osm.org . https://www.openstreetmap.org, 2017.

[Zhang *et al.*, 2017] Zhengxin Zhang, Qingjie Liu, and Yunhong Wang. Road extraction by deep residual u-net. *CoRR*, abs/1711.10684, 2017.

[Zhou *et al.*, 2015] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. *CoRR*, abs/1512.04150, 2015.