

Predicting Road Quality Across Local Governments in Los Angeles County

Andrea Vallebuena
Department of Statistics
Stanford University
avaimar@stanford.edu

Abstract

Road quality is an essential public good for the social and economic inter-connectivity of human establishments. In order to study the dynamics of its provision by local governments, we require fine-grained measurements along a vast number of streets, a highly costly endeavor. For this reason, we turn to street imagery to generate predictions of pavement quality across block groups in the entirety of Los Angeles County, which will be used in future research to study the provision of this public good. We collect a dataset of 151,000 images from the Google Street View Static API and train a Convolutional Neural Network on a 3-class classification task. Image-level predictions are aggregated to produce road quality estimates that are representative of block group-level quality through random sampling of street segments. We explore different model architectures, training methods and loss functions to address challenges including class imbalance stemming from low frequency of low-quality roads and noise in dataset labels. Our best-performing model has a ResNet-50 backbone, and registers a block group-level performance of 78.8% accuracy, 39.1% macro precision and 38.7% macro recall on the test set.

1. Introduction

The maintenance of good quality roads is essential to the social and economic inter-connectivity of human establishments, both at local levels within urban and rural centers, and at global levels connecting centers with other locations of economic activity. As a public good, appropriate pavement quality within urban centers is mostly provided by local governments, and so understanding the political and social factors that dictate this provision is fundamental to societal well-being and equity. We study the case of Los Angeles County to identify how road quality differs across local governments by predicting the pavement quality of street segments belonging to different block groups.

Quantifying pavement quality is a time-intensive and costly endeavor which requires the traversal of each seg-

ment to measure the level of pavement distress that is present in the street. Given that street-level road quality measurements for Los Angeles County are only available for streets located within Los Angeles City and unincorporated county areas, we use Google Street View (GSV) imagery to generate estimates of pavement quality across all block groups in the county which will be used in future research to study the geographic variation in road quality across local governments. Using the labeled street segments in Los Angeles City and unincorporated areas, we build a dataset pairing over 151,000 GSV images with a Pavement Condition Index (PCI) and PCI measurement date, a continuous measure which reflects the density of various pavement distresses along a street segment.

We leverage a bottom-up approach that generates image-level predictions of road quality and aggregates them to generate estimates at the street and block group levels. Our image-level model is a Convolutional Neural Network (CNN) that takes as input a 224 x 224 RGB image of a subset of a street’s pavement, and outputs the predicted road quality class (serious, fair or good), equivalent to a uniformly binned version of the street’s PCI measurement. We experiment with different model architectures to produce segment-level predictions, and randomly sample 10% of the street segments in a given block group to generate representative estimates of road quality at this level. Our best-performing image-level model registers a test-set accuracy of 54.7%, macro precision of 45.1% and macro recall of 51.7%, and its block group-level predictions result in 78.8% accuracy, 39.1% macro precision and 38.7% macro recall on the test set.

2. Literature Review

There is an extensive body of literature using imagery to assess road quality in the form of pavement classification and pavement distress detection.

Road quality classification and measurement. There are several works using street imagery to generate measures of pavement quality. Closest to our task is Ma *et al.* [12], who match GSV imagery with government inspection rat-

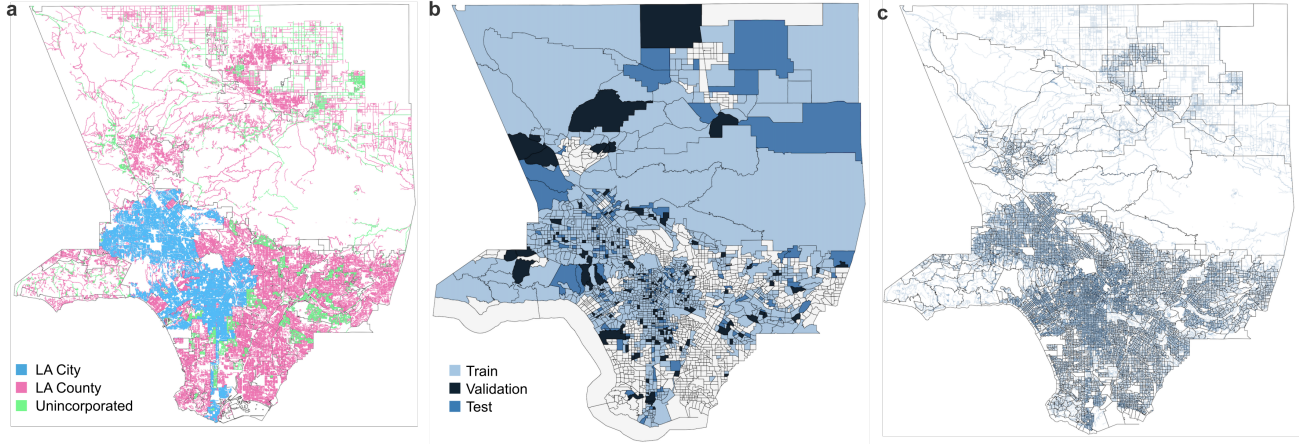


Figure 1. **a)** Network of over 400,000 street segments belonging to Los Angeles City, unincorporated areas and the rest of Los Angeles County. **b)** Census tracts containing labeled street segments from Los Angeles City and unincorporated areas are randomly allocated to training, validation and test splits to build the Training dataset. **c)** Block group-level road quality predictions are computed for 6,572 block groups (black) located in Los Angeles County by sampling 10% of the street segments (blue) in each block group.

ings of street segments in New York City to build a dataset of 700,000 images and train a pavement quality classifier. Classification for a street is performed by generating image-level feature tensors using a VGG-16 network, which are pooled across images to generate an encoding of the segment. This pooling operation is advantageous as it converts an input that varies according to the street length into a fixed-length vector that is fed to a SVM classifier. Also of note is Zhang & Wang [22], who train a CNN which computes the existence of pavement damage by generating pixel-level predictions of the presence of cracks in images of pavement, and Zhang *et al.* [23], who train a binary classifier that predicts the presence of cracking in image patches of pavement. In line with our task, these works assess the presence of cracking regardless of the distress type to quantify general pavement damage, a design choice that simplifies the dataset labeling process.

Road quality measurement has also been explored through the use of satellite imagery. Cadamuro *et al.* [4, 5] pair road quality measurements collected along 57 roads in Kenya, covering 7,000 km in length, with satellite imagery to train a model classifying roads into 5 classes ranging from bad to great road quality. They explore implementing changes in model architecture to capture the sequential nature of road imagery, making use of long short-term memory (LSTM) networks to predict segment-level road quality. Other works involving the extraction of road quality information from satellite imagery are Brewer *et al.* [3], Gebreegziabher [8] and Oshri *et al.* [16]. These discuss essential challenges arising in the setting of road quality prediction from satellite imagery that are also pertinent to a task using street imagery, including the temporal mismatch between imagery and label capture, spatial splitting of the

datasets for appropriate model evaluation, imagery resolution requirements to discern road distresses, and class imbalance due to the relative proportions of lower and higher-quality roads in countries such as the United States.

Pavement distress detection. The pavement distress detection literature focuses on the detection and localization of specific distress classes, such as alligator and longitudinal cracks, rather than the general assessment of pavement quality via classification. Nevertheless, several works are relevant to our task for two reasons. Firstly, they perform an equivalent process of building datasets that pair street imagery with road inspections. Secondly, the types of distress that these works detect are precisely those included in the definition of the PCI. We highlight Maeda *et al.* [13], who collect street imagery from municipalities in Japan and predict 8 types of distresses, and Arya *et al.* [1], who collect street imagery from multiple countries using smart phones and detect 4 damage classes using a Single Shot Detector (SSD) model.

3. Datasets

3.1. Street Segments and PCI

Los Angeles County’s street network is composed of over 400,000 street segments (see Figure 1a). Pavement quality is evaluated every few years for streets belonging to Los Angeles City (over 72,000 segments) and for streets located in unincorporated areas of the county (over 20,000 segments). This evaluation is performed using an automated van that gathers data on the density of various pavement distresses such as alligator cracking and potholes in each street. The PCI of each segment is a real number between 0 and 100 which summarizes the presence of these



Figure 2. **a)** Samples of Serious, Fair and Good-rated street segments. **b), c)** We explore two architectures to generate segment-level predictions of pavement quality, including aggregating image-level PCI predictions (**b)** and training an MLP on image-level extracted 1000-dimensional feature vectors (**c**).

Dataset	Location	Street segments
Training	Los Angeles City	66,705
	Unincorporated areas	17,732
Prediction	Los Angeles County	256,988
	Los Angeles City	66,705
	Unincorporated areas	17,732

Table 1. Training and Prediction street segment datasets

distresses, such that 0 refers to failed and 100 refers to excellent pavement quality. [21].

To train the model, we build a *Training dataset* of street segments located in Los Angeles City and in unincorporated areas that are labeled with a PCI and PCI measurement date, using data from Los Angeles City’s Public Works Department [15]. To predict road quality for the entirety of Los Angeles County, we create a *Prediction dataset* including the segments from the Training set and unlabeled segments from the rest of the county [14]. Given our focus on road quality provision by local governments, we restrict both datasets to include only residential street segments. These datasets are summarized in Table 1.

3.2. GSV

To generate road quality predictions across the entirety of Los Angeles County, we query the Google Street View Static API to obtain a set of images that are representative of the pavement condition of each street segment. We collect a fixed number of 640 x 640 images for each segment such that their locations are randomly sampled along the length of the street. We tune GSV parameters such as field of view (FOV), heading and pitch to generate imagery that is parallel to the direction of the segment and downward-facing such that image coverage of the pavement is maximized. Imagery is collected for each street segment only if it is available within a 9-month window of the recorded PCI measurement date to minimize potential noise in the labels stemming from re-pavement or severe deterioration in pavement quality. The collected 640 x 640 images are pre-processed by cropping a 440 x 440 patch from the lower,

middle part of the image to further maximize the visualization of the street’s pavement and minimize the presence of other objects such as cars and sidewalks. Images are then resized such that the final inputs to the model are 224 x 224 x 3 tensors. Figure 2a) displays the imagery collected for three sample segments corresponding to serious, fair and good PCI measures.

Dataset	Train split	Spatial Coverage	Segments	Images	Images per segment
Training	Train	1,135 census tracts	11,349	120,464	16
	Validation	141 census tracts	1,417	14,864	16
	Test	143 census tracts	1,692	17,712	16
Prediction	-	6,572 block groups	27,716	165,644	8

Table 2. Training and Prediction GSV datasets

In order to conduct fair evaluations of model performance, the training, validation and test splits are realized at the census tract level (see Figure 1b) rather than the segment level, and for each census tract we randomly sample 10% of its street segments to build the Training set. This strategy minimizes potential overlap in the images belonging to each split, as streets are likely to share imagery at their endpoints and intersections. Moreover, this allows for a clearer understanding of the model’s capacity to generalize beyond Los Angeles City by specifically considering performance for census tracts on the outskirts of this urban center. For the prediction set we also randomly sample a fixed percentage of 10% of the street segments in a given block group in order to generate block group-level measures of road quality. Table 2 summarizes the spatial coverage and imagery collected for each dataset and training split.

4. Methods

4.1. Model Architecture

Our approach to perform segment-level pavement quality predictions is the use of a CNN that independently classifies the pavement quality of each image belonging to the segment. We explore two architectures (see Figure 2b, c)

to compute segment-level predictions from this image-level CNN: 1) mean aggregation of the independent image-level classifications; and 2) perform feature extraction on each image, concatenate the extracted features for all images in a segment, and train a multilayer perceptron (MLP) to predict the segment-level road quality class. Road quality predictions at the block group or census tract level are computed by averaging the segment-level predictions of the 10% sampled segments within the administrative unit (census tract or block group).

For the image-level CNN, we employ a ResNet architecture trained on independent image and street label pairs. We note that this image-level training is highly likely to introduce additional noise to the labels given that a single image may not be representative of the average distress density that is captured by the PCI in a given street. This will be further discussed and evaluated in the Results section. We explore both training the ResNet model entirely using Kaiming weight initialization [10], and leveraging model weights pretrained on the ImageNet dataset while freezing a subset of model layers.

Model inputs consist of the 224 x 224 x 3 street images, which are normalized by computing the channel-wise means and standard deviations across the training set in line with standard practice for the ResNet model. Data augmentation is performed via random horizontal flips with a 0.5 probability. Regarding model outputs, although the PCI is a continuous measure we perform the simpler task of classification by binning it into 3 classes: Poor (0-33), Fair (34-66) and Good (67-100). We also explore uniformly binning into 5 classes to understand how sensitive performance is to this discretized formulation of the pavement quality.

To train the CNN, we use an NVIDIA T4 Tensor Core GPU. We build upon the PyTorch [17], PyTorch Lightning [7] and Weights and Biases [2] frameworks for training and training visualization, and implement the pipeline to geocode Los Angeles County (using GeoPandas [11]), clean the PCI datasets, query the GSV imagery, preprocess the inputs, define the model architectures and perform model evaluation from scratch.

4.2. Class Imbalance

The dataset’s severe class imbalance presents a notable challenge, with the training set comprised by Serious (7% of total), Fair (27%) and Good (66%) data points. We make use of both the sampling and model loss approaches that have traditionally been used by the literature to address class imbalance and improve model performance.

In terms of sampling, we make use of the following techniques. Firstly, we explore using weighted re-sampling at training time by inverse class frequencies. Secondly, we build *class-balanced* datasets by over-sampling low-frequency classes and under-sampling high-frequency

Hyperparameter	Values	Baseline model	Best performing model
ResNet size	[18, 34, 50]	18	50
Sampling approach	[None, Weighted sampling, Class-balanced dataset, Two-stage approach]	Weighted sampling	Class-balanced dataset
Number of classes	[3, 5]	3	3
Initial Learning Rate	[1e-1, 5e-2, 5e-3, 8e-3, 1e-4]	1e-4	8e-3
Batch size	[16, 32, 64, 128]	16	128
Weight decay	[0, 1e-1, ..., 1e-5]	0	1e-5
Loss	[Cross entropy, class balanced loss]	Cross entropy	Cross entropy
Layer freezing	[None, Partial]	Partial	None
Weight initialization	[ImageNet, Kaiming]	ImageNet	Kaiming

Table 3. Hyperparameter exploration and hyperparameter values for baseline and best-performing image-level models

classes to generate a dataset of the same size as the imbalanced dataset’s training and validation sets. Thirdly, we use a two-stage training approach to train the model on the imbalanced dataset in the first stage, and train the model’s final linear layer on the class-balanced dataset.

With respect to model loss approaches, we use the class-balanced loss proposed by Cui *et al.* [6], which was designed for long-tailed distribution datasets in which few classes are highly over-represented while a large number of classes contain low sample frequencies. Let n_y be the number of samples in the dataset of class y . Given a vector of predicted class probabilities p and label y , this approach re-weights the loss function for a given data point, $L_i(p, y)$, based on the effective number of samples. The class-balanced loss, $CB_i(p, y)$, is thus the following:

$$CB_i(p, y) = \frac{1}{E_{n_y}} L_i(p, y),$$

where E_{n_y} is the effective number of samples, computed as $E_{n_y} = \frac{1-\beta^{n_y}}{1-\beta}$, where $\beta \in [0, 1)$ is a hyperparameter. We follow the implementation suggestions described by Cui *et al.* to avoid instability when training with this loss, including initializing the biases in the final linear layer to a specific constant based on the number of classes, and setting the initial learning rate to 0.1 while reducing the learning rate decay to 0.01.

4.3. Baseline Model & Hyperparameters

As a baseline model, we use a ResNet-18 architecture for the 3-class classification task with weighted sampling at train time, pretrained ImageNet weights and layer freezing, and cross-entropy loss. Table 3 describes the hyperparameters defining this baseline and outlines the explored hyperparameter values.

5. Results & Discussion

To perform a quantitative evaluation of model performance, we use standard classification metrics including accuracy, and macro and micro precision and recall. Metrics computed on the validation set are used for hyperpa-

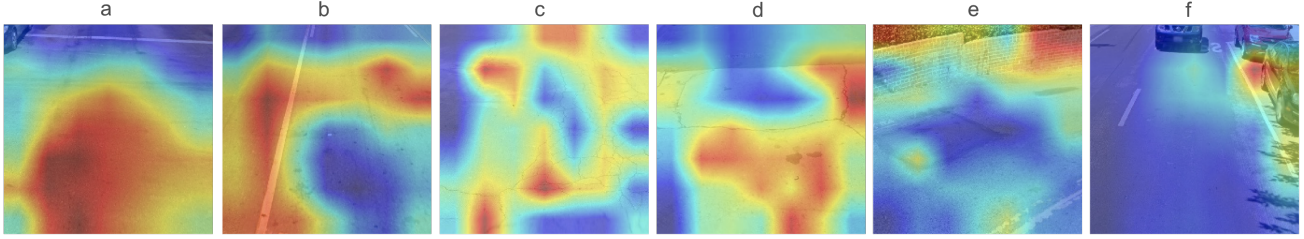


Figure 3. CAMs samples for the best-performing model. **a, b)** Predictions for high road quality images driven by smooth activations across the distress-free pavement; **c, d)** Activations for lower quality roads appropriately driven by the presence of pavement cracking; **e, f)** Predictions remain sensitive to the presence of objects such as cars and shadowing.

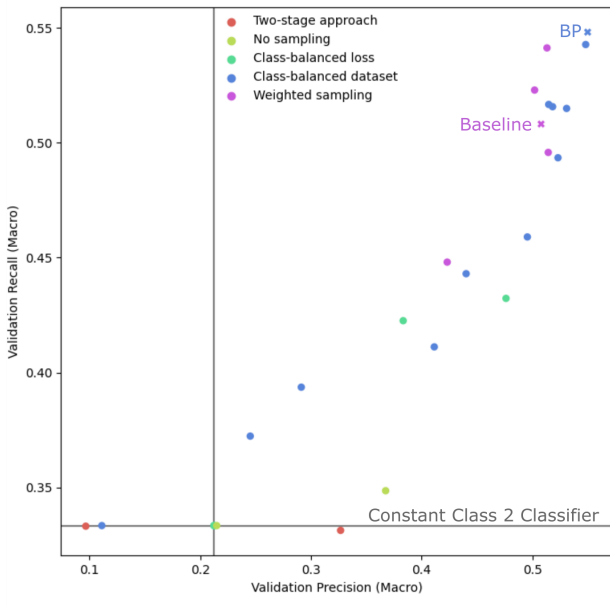


Figure 4. Validation precision (macro) and recall (macro) of image-level models for the 3-class classification task. Horizontal and vertical lines indicate performance of a classifier that consistently predicts the highest frequency class (class 2—good quality). The highlighted baseline and best-performing (BP) models are outlined in Table 3.

parameter tuning, and we report test set metrics for the best-performing model, which is outlined in Table 3. The qualitative results presented throughout this section are also generated using this model.

5.1. Image-level Performance

Figure 4 plots the macro precision and recall on the validation set across the image-level models trained on the 3-class classification task and colored according to the method of training. We use macro metrics, which equally weigh each class, for model selection to ensure solid precision and recall performance across the three classes. The baseline model, a ResNet-18 architecture using pre-trained Imagenet

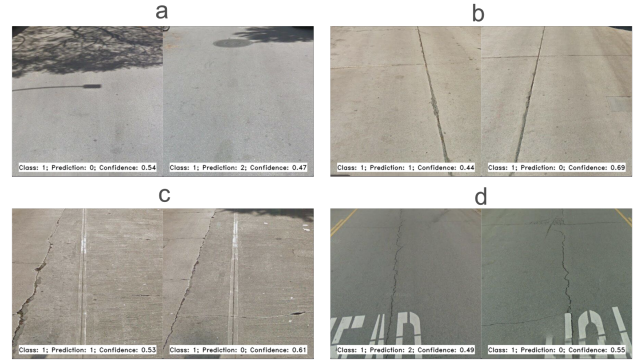


Figure 5. Minor variations in camera angles and capture locations shed light on model prediction drivers, including shadow patterns (a), cracking positions (b, c) and image hues (c).

geNet weights trained using a weighted sampling approach and layer freezing, achieves a macro precision and recall of 50.8% and 50.8%, respectively. The best-performing model is a ResNet-50 trained on the class-balanced dataset (see Table 3 for other hyperparameters), and registers macro precision and recall of 55.0% and 54.8%, respectively (accuracy of 54.7%, macro precision of 45.1% and macro recall of 51.7% on the test set). These figures compare to macro precision and recall of 21.2% and 33.3% for a classifier which consistently predicts the highest frequency class. The best-performing image-level model trained on the 5-class task achieves a macro precision and recall of 31.8% and 31.9%, respectively. We expect that this closer performance of the 5-class model to that of a random classifier is partially driven by increased noise in the labels stemming from discretizing the PCI into a higher number of classes. Please refer to section 6 of the Supplementary Materials for a comprehensive comparison of validation accuracy, precision and recall across all experiments.

To evaluate model performance we also perform the following qualitative analyses to examine the physical attributes in the images that are driving model predictions. Firstly, we visualize Class Activation Maps (CAMs) of the

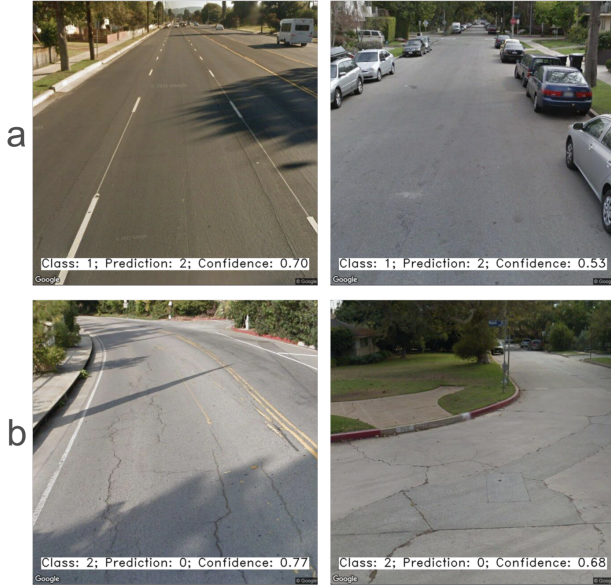


Figure 6. The image-level dataset suffers from noisy data point labels, occasionally resulting in high confidence, correct predictions on mislabeled images. Mislabeled high-quality (a) and low-quality (b) streets with correct model predictions.

images (see Figure 3), which reveal that the model’s predictions are generally driven by the absence of distresses across the pavement (Figure 3 a, b)) and by the presence of pavement cracking in low quality roads (Figure 3 c, d)). However, these maps also reveal that predictions can be sensitive to the presence of objects such as cars and fences, and shadows (Figure 3 e, f)).

Secondly, a deeper understanding of model predictions can be achieved by leveraging instances in which the collected imagery captures a specific location and viewpoint, with minor shifts in camera angles or distance. Should model predictions or confidence in predictions vary on a pair of images of this nature, we can attribute the variation in the images as a driver for this change in model output. GSV imagery naturally allows for these comparisons, as it is common for different visualizations of the same location to be available via minor changes in the geocoded location or in GSV querying parameters. Figure 5 visualizes some image pair examples of this nature, revealing that predictions for the same patch of a street can vary with the presence of different shadow patterns (Figure 5a)), closeness and angles of pavement cracking (Figure 5b, c)), and potentially the image hue (Figure 5d)). This suggests that improved model performance could be achieved through a more aggressive form of data augmentation through randomly cropped image patches.

As discussed in section 4.1, performance metrics are not wholly reflective of the model’s capacity to predict road

quality given the image-level noise in the labels and potential discrepancies between the PCI measurement date and the date of GSV imagery capture. We perform error analysis on the model’s test set predictions to examine whether this issue is driving an incorrect assessment of model performance, and in particular examine misclassified, potentially mislabeled data points for which the model makes predictions with a high level of confidence. Figure 6 visualizes samples in which ground truth road quality is good but has been labeled as fair, and the model has successfully predicted the ground truth label (6a). Conversely, 6b displays samples labeled as having good pavement quality, which the model has correctly predicted with high confidence to be of serious quality.

5.2. Segment-level Performance

As discussed in the section 4.1, we experiment both using mean aggregation of image-level predictions¹ and feature extraction to compute segment-level predictions. The best-performing segment-level model, as measured in terms of macro precision and recall, employs the feature extraction and MLP architecture. Figure 7 visualizes the ground truth road quality labels (7a), the predictions generated by this model (7b) and the prediction errors (7c) for each segment in the training set. As observed in 7b, c model predictions tend to overestimate road quality compared to the ground truth and for this reason lack some of the granularity exhibited in road quality patterns by the ground truth labels, especially within Los Angeles City. No discernible geospatial patterns are apparent regarding consistent overestimation or underestimation of segments within specific regions, which could be cause for concern when the predictions are used as inputs downstream in a statistical inference setting.

5.3. Block group-level Performance

Road quality predictions at the block group-level are performed via mean aggregation of the 10% of segments sampled for the block group, using the segment-level model presented in section 5.2. Figure 8 plots the absolute error in block group-level road quality for block groups in the training (8a), validation (8b) and test (8c) sets. Across each training split, it can be observed that the block group-level predicted road quality classes rarely have errors that are more than 1 class apart from the ground truth labels. Limited model overfitting is evidenced by comparable model performance in the test set and the validation set, standing at 78.8% accuracy, 39.1% macro precision and 38.7% macro recall in the test set compared to 69.0% accuracy, 39.1% macro precision and 37.9% macro recall in the validation

¹For computational reasons, we use an image-level model that is smaller than the best-performing model presented in section 5.1, but has comparable performance, to generate the segment-level and administrative unit-level results discussed in sections 5.2 and 5.3.

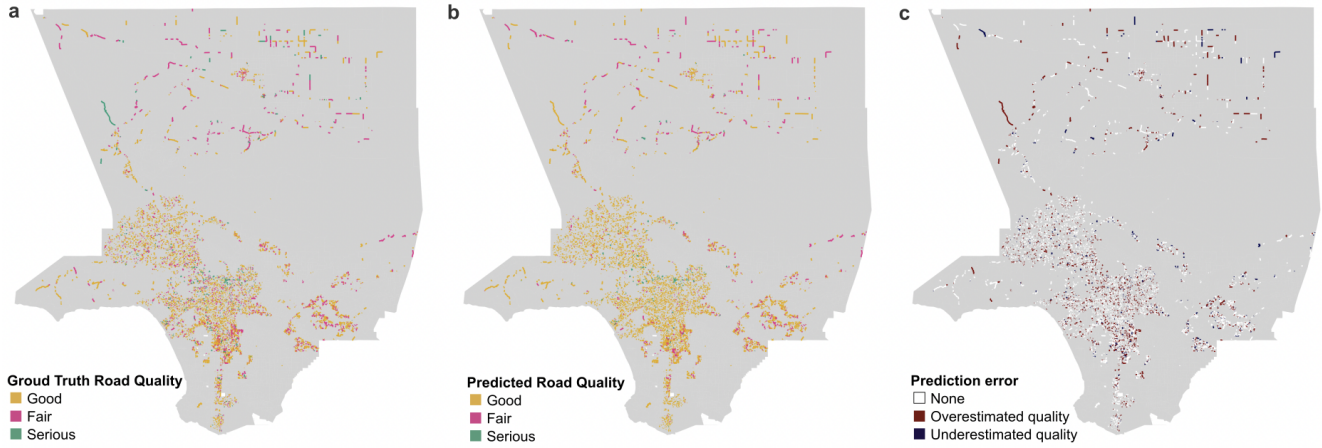


Figure 7. Segment-level road quality ground truth (a), model predictions (b) and prediction errors (c) in the GSV Training dataset covering Los Angeles City and unincorporated areas of Los Angeles County.

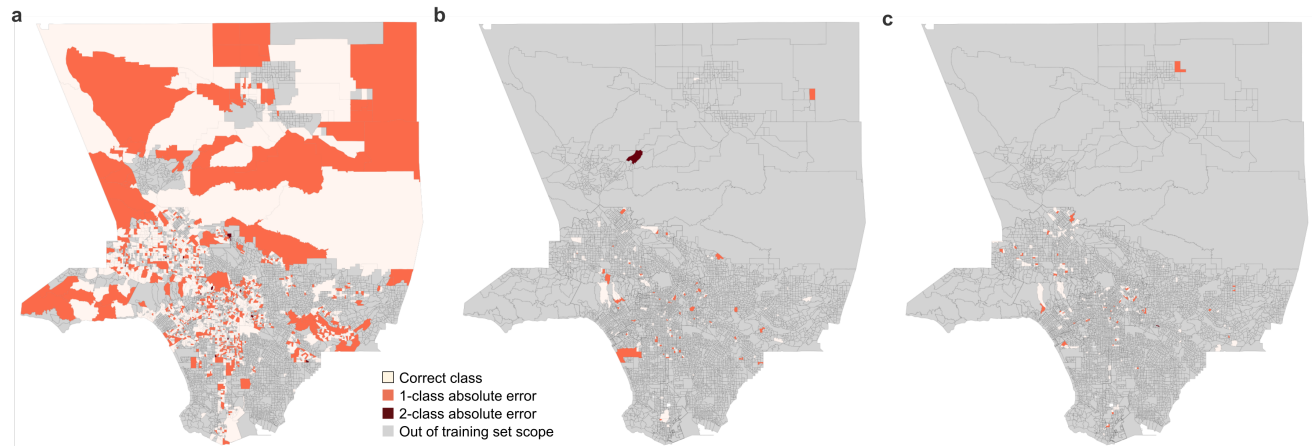


Figure 8. Block group-level road quality predictions for block groups in the training (a), validation (b) and test (c) sets.

set. In line with the segment-level analysis, error dynamics do not present any discernible geographic patterns.

In order to correctly evaluate model predictions at the block group level, we also distinguish between prediction error and segment sampling error, given that we seek to predict block group-level road quality with a representative sample of segments. Computing predictions using the ground-truth labels of the 10% sampled segments as opposed to the model predictions for this subset of segments highlights relevant limitations in the model's capacity to assess road quality at the administrative level stemming from the segment sampling design rather than from the segment-level classifier's performance. Accuracy, macro precision and macro recall in the test set stand at a close 78.4%, 44.5% and 49.6% using the ground truth labels for the sampled segments (compared to 78.8%, 39.1% and 38.7% using model predictions). This underscores the opportunity to potentially improve administrative unit-level performance by

increasing the percentage of sampled segments and reducing the imagery captured per segment.

A potential concern is whether the urban setting within Los Angeles City, where the majority of the streets in the training set are located, differs from that of other parts of Los Angeles County and could cause the model to generalize poorly to other locations within the county. To this end, we compare performance metrics in the test set for block groups within (218 block groups) and outside (23 block groups) Los Angeles City. While accuracy presents a notable gap across these geographies (80.7% within Los Angeles City vs. 60.9% in the rest of the county) driven by higher quality roads within the city (mean PCI is 76 vs. 69 in the rest of the county), macro precision and recall display not only comparable but improved performance outside of Los Angeles City (40.7% precision and 38.2% recall vs. 38.3% and 38.9%, respectively).

6. Conclusion

We have designed and trained a series of model architectures to generate predictions of serious, fair and good pavement quality at the image, street and block group levels based on a dataset of GSV imagery and segment-level measurements of road quality in Los Angeles City and unincorporated areas of Los Angeles County. Our main objective was to produce good quality predictions of pavement quality at the block group level across the entirety of Los Angeles County, to be used downstream as inputs in a statistical inference setting exploring the relationship between local governments and the provision of public goods.

We executed a bottom-up approach, which was trained on and ran inference at the level of a single image, and aggregated predictions to produce estimates for streets and block groups. The highest performing model was a ResNet-50, which benefited from end-to-end training on a balanced class dataset. This outperformed other methods used to combat the class imbalance that was present in the dataset, including weighted sampling, the use of class-balanced loss and two-stage training.

In future work we primarily aim to address two challenges that we found to constrain model performance throughout this work. The first is the noise in the image-level road quality labels that is present due to mismatches between the PCI measurement date and the images capture dates and due to the spatial distribution of pavement distresses. This issue can be partly corrected through several methods. Firstly, a trained image-level model with solid performance can be used to re-label the dataset for data points with high confidence predictions, thus improving model performance in an iterative fashion. Secondly, a change in architecture such that the model is trained end-to-end at a level that matches the unit of labeling could be beneficial to correct this mismatch. An example of this kind of architecture is a Convolutional LSTM network, which could receive various images along a segment as input to directly predict segment-level road quality. The second challenge is the severe class imbalance exhibited by the collected imagery. In this case, a more efficient imagery collection process for the training and validation sets which considers the creation of a balanced dataset at the segment sampling stage could provide sufficient data to ensure solid performance across all classes.

Contributions & Acknowledgements

We wish to acknowledge [David Schönholzer](#), who defined the research question driving the inception of this project, offered insightful discussions throughout its execution and provided the funding for the GSV imagery. We also wish to thank Stanford's CS 231N team, in particular Agrim Gupta, for providing valuable guidance.

Code Usage. We build upon the PyTorch [17], PyTorch Lightning [7], Weights and Biases [2] and GeoPandas [11] frameworks, libraries and tutorials to build our code base. We also leverage third-party code to generate the CAMs visualizations [9], and query the GSV imagery [18], and referenced the following PyTorch implementations of the class-balanced loss ([20] and [19]).

Supplementary Materials

Experiments

See Table 4.

References

- [1] Deeksha Arya, Hiroya Maeda, Sanjay Kumar Ghosh, Durga Toshniwal, Alexander Mraz, Takehiro Kashiyaama, and Yoshihide Sekimoto. Deep learning-based road damage detection and classification for multiple countries. *Automation in Construction*, 132:103935, 2021. 2
- [2] Lukas Biewald. Experiment tracking with weights and biases, 2020. Software available from wandb.com. 4, 8
- [3] Ethan Brewer, Jason Lin, Peter Kemper, John Hennin, and Dan Runfola. Predicting road quality using high resolution satellite imagery: A transfer learning approach. *PLOS ONE*, 16(7):1–18, 07 2021. 2
- [4] Gabriel Cadamuro, Aggrey Muhebwa, and Jay Taneja. Assigning a grade: Accurate measurement of road quality using satellite imagery. *CoRR*, abs/1812.01699, 2018. 2
- [5] 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 SIG-CAS Conference on Computing and Sustainable Societies, COMPASS '19*, page 145–154, New York, NY, USA, 2019. Association for Computing Machinery. 2
- [6] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge J. Belongie. Class-balanced loss based on effective number of samples. *CoRR*, abs/1901.05555, 2019. 4
- [7] William Falcon et al. Pytorch lightning. *GitHub. Note: <https://github.com/PyTorchLightning/pytorch-lightning>*, 3, 2019. 4, 8
- [8] Bisrat Araya Gebreegziabher. Quality from optical satellite imagery using machine learning. 2
- [9] Jacob Gildenblat and contributors. Pytorch library for cam methods. <https://github.com/jacobgil/pytorch-grad-cam>, 2021. 8
- [10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. 4
- [11] Kelsey Jordahl, Joris Van den Bossche, Martin Fleischmann, Jacob Wasserman, James McBride, Jeffrey Gerard, Jeff Tratner, Matthew Perry, Adrian Garcia Badaracco, Carson Farmer, Geir Arne Hjelle, Alan D. Snow, Micah Cochran, Sean Gillies, Lucas Culbertson, Matt Bartos, Nick Eubank, maxalbert, Aleksey Bilogur, Sergio Rey, Christopher Ren, Dani Arribas-Bel, Leah Wasser, Levi John Wolf, Martin Journois, Joshua Wilson, Adam Greenhall, Chris Holdgraf,

ResNet Size	Frozen Layers	Batch size	Initial learning rate	Training method	Weight decay	Classification task	Accuracy (train)	Accuracy (validation)	Macro precision (validation)	Macro recall (validation)
50	0	128	8e-3	No sampling	1e-5	3 class	0.66	0.64	0.37	0.35
50	0	128	1e-1	Class-balanced loss	1e-5	3 class	0.51	0.57	0.38	0.42
50	0	128	8e-3	Class-balanced dataset	1e-5	3 class	0.59	0.55	0.55	0.55
50	0	128	8e-3	Class-balanced dataset	1e-4	3 class	0.41	0.39	0.29	0.39
50	0	128	8e-3	Class-balanced dataset	1e-3	3 class	0.40	0.37	0.25	0.37
34	0	16	1e-1	Class-balanced loss	1e-2	3 class	0.57	0.64	0.21	0.33
50	165	128	5e-5	Two-stage approach	1e-5	3 class	0.36	0.32	0.33	0.32
34	114	64	1e-3	Two-stage approach	0.00	3 class	0.34	0.08	0.10	0.33
50	0	64	8e-3	Class-balanced dataset	0.00	3 class	0.75	0.49	0.52	0.49
34	0	64	5e-2	Class-balanced loss	0.00	3 class	0.67	0.64	0.48	0.43
34	0	64	8e-3	Class-balanced dataset	0.00	3 class	0.60	0.51	0.53	0.51
34	0	128	5e-2	Class-balanced dataset	0.00	3 class	0.58	0.52	0.52	0.52
34	0	128	1e-2	Class-balanced dataset	0.00	5 class	0.41	0.32	0.32	0.32
34	0	64	1e-2	Class-balanced loss	0.00	3 class	0.66	0.64	0.21	0.33
34	0	64	1e-2	Class-balanced dataset	0.00	5 class	0.45	0.29	0.28	0.29
34	0	256	2e-2	Class-balanced dataset	0.00	3 class	0.49	0.44	0.44	0.44
34	0	128	1e-2	Class-balanced dataset	0.00	3 class	0.56	0.52	0.51	0.52
34	0	64	1e-2	Class-balanced dataset	0.00	3 class	0.64	0.54	0.55	0.54
34	0	16	1e-2	Class-balanced dataset	0.00	3 class	0.33	0.33	0.11	0.33
34	0	32	1e-2	Class-balanced dataset	0.00	3 class	0.88	0.46	0.50	0.46
34	0	32	4e-2	Class-balanced dataset	0.00	3 class	0.41	0.41	0.41	0.41
34	0	32	4e-2	Weighted sampling	0.00	3 class	0.61	0.52	0.42	0.45
34	0	32	4e-2	No sampling	0.00	3 class	0.56	0.64	0.21	0.33
34	0	32	4e-2	No sampling	0.00	5 class	0.41	0.39	0.08	0.20
18	30	16	2e-2	Weighted sampling	0.00	3 class	0.78	0.60	0.51	0.54
18	45	16	8e-3	Weighted sampling	0.00	3 class	0.83	0.59	0.50	0.52
18	45	32	5e-3	Weighted sampling	0.00	3 class	0.86	0.61	0.51	0.51
18	45	16	5e-3	Weighted sampling	0.00	3 class	0.84	0.64	0.51	0.50

Table 4. Accuracy, macro precision and macro recall on the validation set across experiments.

- Filipe, and François Leblanc. geopandas/geopandas: v0.8.1, July 2020. 4, 8
- [12] Ke Ma, Minh Hoai, and Dimitris Samaras. Large-scale continual road inspection: Visual infrastructure assessment in the wild. In *Proceedings of British Machine Vision Conference*, 2017. 1
- [13] Hiroya Maeda, Yoshihide Sekimoto, Toshikazu Seto, Takehiro Kashiya, and Hiroshi Omata. Road damage detection using deep neural networks with images captured through a smartphone. *CoRR*, abs/1801.09454, 2018. 2
- [14] County of Los Angeles. The Los Angeles County County-wide Address Management System, 2022. 3
- [15] Department of Public Works City of Los Angeles. Road Surface Condition Map, 2020. 3
- [16] Barak Oshri, Annie Hu, Peter Adelson, Xiao Chen, Pascaline Dupas, Jeremy Weinstein, Marshall Burke, David Lobell, and Stefano Ermon. Infrastructure quality assessment in africa using satellite imagery and deep learning. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’18, page 616–625, New York, NY, USA, 2018. Association for Computing Machinery. 2
- [17] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimsheine, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. pages 8024–8035, 2019. 4, 8
- [18] robolyst. streetview. <https://github.com/robolyst/streetview>, 2021. 8
- [19] statsu1990. Yoto class balanced loss. https://github.com/statsu1990/yoto_class_balanced_loss, 2021. 8
- [20] vandit15. Class balanced loss pytorch. <https://github.com/vandit15/Class-balanced-loss-pytorch>, 2021. 8
- [21] Los Angeles County Public Works. Pavement Management System, 2022. 3
- [22] Allen Zhang, Kelvin C. P. Wang, Baoxian Li, Enhui Yang, Xianxing Dai, Yi Peng, Yue Fei, Yang Liu, Joshua Q. Li, and Cheng Chen. Automated pixel-level pavement crack detection on 3d asphalt surfaces using a deep-learning network. *Computer-Aided Civil and Infrastructure Engineering*, 32(10):805–819, 2017. 2
- [23] Lei Zhang, Fan Yang, Yimin Daniel Zhang, and Ying Julie Zhu. Road crack detection using deep convolutional neural network. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3708–3712, 2016. 2