Classifying Road Quality Using Satellite Imagery in Metro Detroit

Nabil Ahmed
nabilah@stanford.edu

Trent McMullen
trentmcm@stanford.edu

David Karamardian
dk11@stanford.edu

Abstract

Our project utilizes transfer learning to classify the quality of roads in Metro Detroit. We cultivate a new dataset of 40,000 satellite images queried from the Google Static Maps API and combine them with publicly available PASER road evaluation data. After addressing significant levels of noise in the PASER road quality labels, we experiment with a pretrained version of ResNet-50 and different loss functions to achieve over 70% classification accuracy on our validation set. Finally, we visualize our model results and feature embeddings through the use of saliency maps and t-SNE cluster analysis.

1. Introduction

Infrastructure deterioration is an issue that plagues much of the United States, with the problem being particularly acute in post-industrial cities like Detroit [7]. Currently, the status of roadways in the city is assessed by municipal workers physically driving the roads and using visual inspection to rate road quality as either "poor", "fair", or "good" according to the Pavement Surface Evaluation and Rating System (PASER). Due to budget constraints, the frequency and scope of PASER updates in Metro Detroit has been reduced, with the process being skipped in large areas of the city in past years. This hampers the ability of local governments to understand their infrastructure needs and allocate resources appropriately.

In a bid to create a more automated and efficient road evaluation process, we have designed and trained a model that utilizes satellite imagery to match human classification of roads based on their quality. We have built a dataset to train this model using publicly available PASER data from 2017-2019 and satellite imagery from the Google Maps Static API, and we leverage a transfer learning approach using ResNet-50. In addition, our model is tuned to reflect the ordinal nature of the target categories and the real-world implications of allocating resources based on its predictions.

2. Related Work

The use of satellite and aerial imagery for computer vision tasks has been an active and fruitful area of research. Prior work exists that focuses on image segmentation to generate overhead maps of roads in large urban areas [8]. Satellite imagery tends to lend itself nicely to machine learning tasks due to qualities such as large dataset size, fixed scale, and common perspective. However, the noisiness of image labeling remains a challenge. In addition, satellite images of roads can suffer from complications such as image occlusion and changes due to seasonal or day/night cycles. To mitigate these problems, data augmentation [4] and use of image patches has been employed.

The particular application of assessing road quality has also benefitted and suffered from the inherent traits of aerial imagery. The main issue is appropriate labeling of road condition, which some researchers have attempted to solve with a combination of sensors attached to vehicles and manual human labeling [1]. Such work has also leveraged existing models such as ResNet-50 via transfer learning for image classification, with the intuition that the pre-trained model has already learned to identify distinctive shapes and features of importance in image classification tasks [6]. Our project expands on this transfer learning approach.

The term "road quality" also deserves some further clarification. Prior research has explored using satellite imagery to assess the quality of infrastructure in developing countries, with the distinction of "poor" versus "good" quality mainly focusing on the presence of hard pavement to indicate a robust roadway [2]. Contextual clues regarding the urbanization of the surrounding environment also aided in road classification. The authors of this work were able to achieve good results under this broader definition of road quality and achieved above 70% classification accuracy on a 5-category task. However, there is greater complexity when applying a similar approach to the Metro Detroit area, as the PASER system we utilize has been calibrated for a fully developed and paved road network. The features that distinguish a "poor" versus "good" road will be more nuanced and potentially more difficult for a model to perceive due to this refinement in the classification objective. While accuracy metrics from these developing infrastructure-type stud-

ies can serve as a baseline, it is important to remain cognizant of this key difference.

3. Data

Our dataset was self-created by combining PASER road evaluation data published by the Southeast Michigan Council of Governments (SEMCOG) with satellite images pulled from the Google Static Maps API. We were thus able to obtain a large quantity of labeled images covering a diverse set of roadways throughout Metro Detroit while only using free data resources.

3.1. PASER Road Condition Dataset

Road quality ratings from 2017-2019 for Metro Detroit were obtained from the Southeast Michigan Council of Governments (SEMCOG) Open Data Portal [12]. The raw 1-10 ratings were produced by expert human evaluators following the PASER system, which focuses on visual inspection of road segments [14]. Each distinct observation in this set is a quality evaluation for a road segment that typically varies in length between 0.01 and 0.50 miles and includes highways, county roads, and residential streets. While the PASER numerical rating for the segment is assigned based on the precise type of deformities visible in the road (longitudinal cracks, transverse cracks, "rutting"), these features can be difficult to distinguish in a satellite image. To simplify the classification task, the ratings were grouped into poor (rating of 1-3), fair (4-7), and good (8-10) to provide for 3 quality levels that can be evaluated from aerial imagery. The initial data gathered was a weighted random sample of 40,000 ratings available in the PASER set for the desired time range, with less weight given to frequent road types like two-lane county roads. Each road segment carries a Beginning Mile Point (BMP) and Ending Mile Point (EMP), which were mapped to GPS latitude and longitude coordinates via ArcGIS shapefiles.

3.2. Google Static Maps API Images

Following the collection of the PASER data, GPS coordinate pairs for each road evaluation were sent to the Google Static Maps API to obtain satellite images. The API allows for requests such as the removal of any road labeling and capturing the image with a zoom level sufficient to see features such as individual vehicles, road markings, and road deformities. Each image returned maintained a dimension of 480x480 pixels, but slight variations in scale and perspective meant that translation to actual distance on the ground fluctuated around 600 by 600 feet. Example images and their corresponding PASER road quality ratings are shown in Figure 1.

(a) Good Quality Road: New, Undamaged Pavement

(b) Fair Quality Road: Hairline and Block Cracking

(c) Poor Quality Road: Severe Cracking, Structural Damage

Figure 1: Example Satellite Images in Dataset

3.3. Data Challenges

A cursory inspection of the initial data gathered revealed a disconnect in many images between their apparent road quality and the road quality rating assigned by the PASER data. Significant levels of such label noise would make training any classification model on the dataset extremely difficult, as a large number of images with visible road deformities were tagged in PASER as "good" roads, while other seemingly unblemished roads were marked as "poor."

To quantify the level of noise in the PASER road condition labels, a human expert in the area of PASER evaluation was consulted to provide their rating on 150 randomly selected images later allocated to the test set. These ratings are compared against the labels available in the PASER road condition dataset in Figure 2.

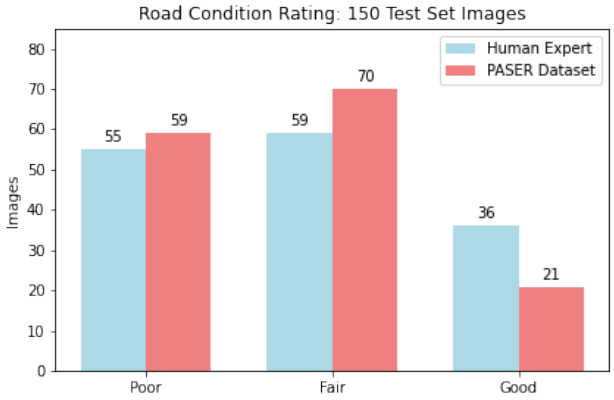

Figure 2: Human Expert Rating vs. PASER Dataset Label

Unfortunately, the human expert only agreed with the PASER label on 58.7% of the images examined. Furthermore, as seen in Figure 2, the human expert was more likely to assign a higher quality rating as compared to the PASER label. An investigation of the discrepancies revealed the following list of causes:

1. **Time Gap Between PASER and Photo Capture**
   There can be a multiple years-long gap between the date that human inspectors assigned the PASER rating for a particular road segment and the date of satellite photo capture for the image returned by the Google Static Maps API, during which time road conditions can change significantly.

2. **GPS Coordinates Uncentered on Road**
   The GPS coordinate pair used to return a satellite image for a given road segment was an interpolation between the beginning and ending mile points, but this could lead to an image that was not centered on critical road features or missed the road entirely.

(a) Satellite Image Not Centered on Road

(b) Human Expert Label: Good, PASER Label: Poor

Figure 3: Problematic Images in Initial Dataset

3. **Low Resolution Images**
   Our human expert noted that some types of road deformities, including underlying structural damage, could not be accurately assessed in the resolution provided by Google Static Maps.

Examples of images driving the label noise are provided in Figure 3.

3.4. Hand Labeling

To address the noise in the PASER data labels, our team hand-assigned labels to 2,250 satellite images for training and 450 for validation, following the guidelines outlined by the PASER system manual [14] and our human expert. Fur-

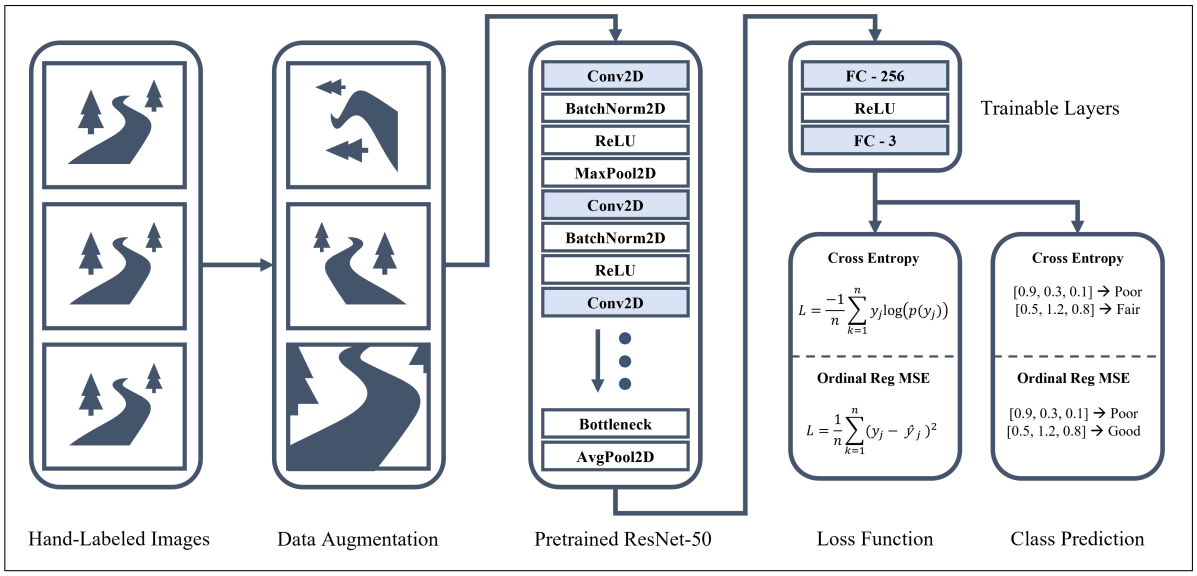

Figure 4: Transfer Learning Based Model Training and Inference Pipeline

thermore, this hand-crafted dataset was built to include a diverse set of road types and contained an even proportion of poor, fair, and good quality roads. For the test set, 5,000 images were randomly sampled from the set that was not allocated for training or validation. As these test images were not hand-screened and therefore only contain the noisy PASER labels, they were not suited for evaluating classification accuracy; rather, they were set aside for analysis and visualization of feature embeddings generated by our model. Table 1 summarizes the final dataset used for training, evaluation, and visualization after the aforementioned hand-labeling procedure.

Split	Poor	Fair	Good	Total
Train	750	750	750	2250
Validation	150	150	150	450
Test	2164	1886	950	5000

Table 1: Images in Dataset for Training / Evaluation

4. Road Quality Classification Task

Transfer learning using a version of ResNet-50 pretrained on ImageNet [5] data was critical to the success of our classification model. Data augmentation steps complementary to ResNet-50 and appropriately selected loss functions also helped to achieve high classification accuracy on both the training and validation sets.

4.1. Data Augmentation

While the hand-labeled dataset contained images that were characteristic of their respective road quality categories, it suffered from limited size. Accordingly, data augmentation was employed to enhance the number of different images seen by the model during training and help it to learn only relevant features. This technique has been employed with great success in past research on image classification problems [4].This augmentation including center-cropping images to focus the image only on the road, random flips and rotations to reflect the diversity of road orientations, and normalizing images to prepare them for ResNet-50:

- Training Transformations
  - Center Crop 256x256
  - Resize 224x224
  - Random 15 Degree Rotation
  - Random Horizontal Flip
  - Normalize with ImageNet Mean and Variance
- Validation Transformations
  - Center Crop 256x256
  - Resize 224x224
  - Normalize with ImageNet Mean and Variance

4.2. Model Architecture

The overall data pipeline and model architecture is described in Figure 4. After data augmentation, the images

were sent through a version of ResNet-50 pretrained on ImageNet to extract meaningful feature representations. The core of ResNet-50 involves repeated layers of convolutional filters, batch normalization, ReLU activations, and max pooling, and the pretrained version of this model has the weights associated with these layers already set to pull useful information from input images [6]. In our transfer learning approach, these layers and weights were left unmodified throughout training. On the other hand, the final fully-connected layers of ResNet-50 were modified to tune for our specific classification task by including a 256-neuron layer, ReLU, and the ultimate 3-neuron output layer containing the class scores. These final fully connected layers were trained using one of the following two loss functions, each with its own advantages:

1. **Target One-Hot Encoding with Cross-Entropy Loss** The classic approach for classification tasks is to embed the target class using one-hot encoding, with the output of the neural network interpreted as a probability of the image lying in the given class. While this method is standard, it does not reflect the fact that our categories are ordered.

2. **Target Ordinal Regression Labels with MSE Loss** This approach provides a way to encourage the model to reflect the ordinal relationship in the class predictions by expressing the target label as a vector specified below, with loss calculated as the mean-squared error between the model output and the target. Prediction vectors can be read left to right and interpreted as the probability of the image being greater than the class at the corresponding index [3]:

$$\text{Poor} \rightarrow [1,0,0] \quad \text{Fair} \rightarrow [1,1,0] \quad \text{Good} \rightarrow [1,1,1]$$

4.3. Model Training

The data pipeline for model training was implemented in PyTorch [9] following the design in Figure 4. The model was trained separately with the cross-entropy loss framework (CE) and ordinal regression loss framework (ORD) to determine which was superior for the classification task. Each training cycle was run for 10 epochs and utilized the ADAM optimizer with the learning rate set to $5 * 10^{-4}$. This combination of optimizer and learning rate were settled upon after a few rounds of trial and error. The evolution of training and validation results for each loss function are presented in Figure 5. Note that the loss functions provide values on different scales.

While the ordinal regression loss version of the model initially learned more quickly, cross-entropy provided a slightly higher final training and validation accuracy.

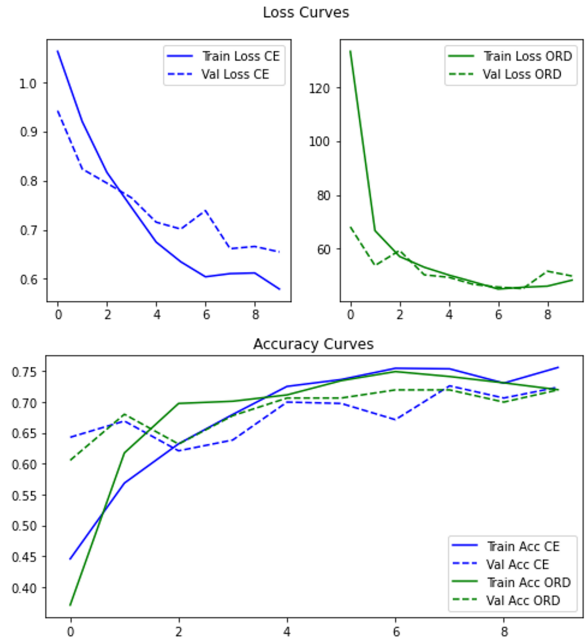

Figure 5: Training Loss and Accuracy vs. Epochs

Model	Training	Validation
ResNet-50 CE	75.6%	72.4%
ResNet-50 ORD	72.0%	71.9%

Table 2: Overall Classification Accuracy

4.4. Results

Both models were able to achieve over 70% classification accuracy on the training and validation sets, with the cross-entropy model hitting slightly higher metrics as shown in Table 2.

However, it is also important to consider the nature of the misclassifications–it is preferable for a poor road to be misclassified as fair rather than good, and there may be greater societal cost for misclassifying a damaged road over an adequate one. As can be seen in Figure 6, the ordinal regression model performs better for truly poor quality roads. This means that for a model application such as identifying roads in need of immediate repair, the ordinal regression model would be appropriate to use despite having slightly lower overall classification accuracy. With these other types of criteria in mind, the ordinal regression model makes a compelling case for its superiority.

Another tool for understanding the model results are saliency maps. These maps identify the pixels in an image that have the greatest influence over the model's class prediction via backpropagation, and thus allow for human interpretation of which features the model has learned to as-

		ResNet-50 CE Predicted			ResNet-50 ORD Predicted		
		Poor	Fair	Good	Poor	Fair	Good
Actual (Hand Label)	Poor	27.2%	5.0%	0.8%	31.4%	1.8%	0.0%
	Fair	6.1%	17.5%	9.6%	10.5%	18.2%	4.6%
	Good	1.3%	4.6%	27.6%	3.5%	7.7%	22.4%

Figure 6: Validation Set Classification Matrix

sociate with each class [11]. Figure 7 provides an example of a saliency map with bright red dots plotted on the parts of the image with the greatest influence over the model's (correct) classification of the road as good quality. The red dots are arranged in a manner that follows the lanes of the road and skips over road markings, meaning that the model has correctly learned to associate continuous and unbroken black or gray lanes with high-quality roads. Road markings are ignored, likely because the model has seen high-quality residential roads that do not contain any of them. However, we can also see that model places some amount of emphasis on the sidewalk in the lower-left hand side of the image, indicating that surrounding context still governs some of the model's prediction behavior.

Figure 7: Saliency: Good Road, Predicted Good (CE)

Figure 8 provides an example of a saliency map for a misclassified image in the validation set. Once again, the model placed attention on surrounding features such as the road shoulder and a side connecting road, which were in relatively good condition. These confounding attributes appear to be the driver of the misclassification.

5. Test Set Feature Embedding and t-SNE Cluster Analysis

Due to the noise in the labels in the test set, any calculated metrics for it related to classification accuracy are unlikely to be very meaningful. Instead, we evaluate how our models generalize to the test data by qualitative analysis of the feature embeddings they extract from images through the use of t-SNE cluster visualizations.

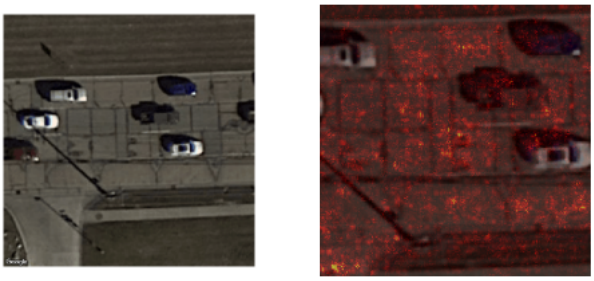

Figure 8: Saliency: Poor Road, Predicted Fair (CE)

5.1. t-SNE Visualization

During inference using our model, the penultimate layer contains a 256-dimension vector representation of a given input image. Intuitively, we would expect to see a lower arbitrarily defined "distance" between road images in the same quality category as compared to roads of different quality. Moreover, the ordinal nature of our categories means that the distance between poor and good roads should be greater than poor to fair roads. t-Distributed Stochastic Neighbor Embedding provides a way to visualize the closeness of the high-dimensional vector representations produced by our model by reducing the vectors into a lower-dimensional space that preserves small pairwise distances. [13]

Using the training set, we produced the two-dimensional t-SNE embedding for each image using both the CE and ORD models and assigned a color according to the handcrafted quality label. As seen in Figures 9 and 10, the plots for both models resulted in nicely shaped clusters between the quality categories. Moreover, the clusters corresponding to fair quality roads sit in between the poor and good clusters (more so for ORD than CE), indicating the models' feature representations are properly reflecting the ordinality of the labels for the training data. While there is noticeable overlap between the clusters, this is to be expected, as road quality lies on an inherently continuous spectrum, with our 3 assigned categories themselves being a simplification of the 1 through 10 PASER rating scale.

Cluster analysis was also performed on the feature embeddings produced for test set images. Due to the significant noise in the PASER labels, the data points were colored according to their predicted category as a proxy for a "true" quality label. Once again, the models both produced distinct clusters, indicating that they were able to generalize their classification power to the test set. However, it should be noted that the clusters for the CE model did not maintain a sequential relationship as it had exhibited for the training

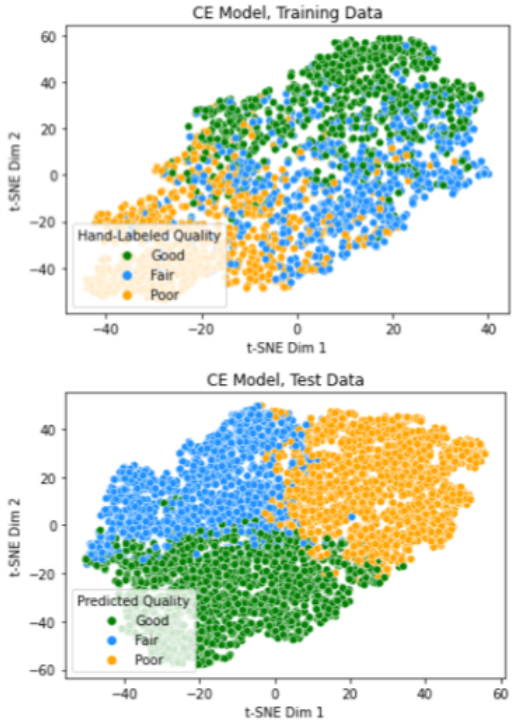

Figure 9: t-SNE Plots for Cross-Entropy Model

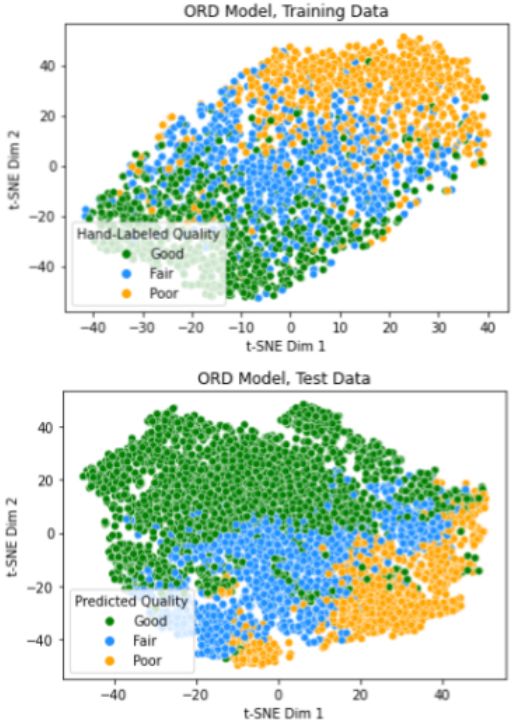

Figure 10: t-SNE Plots for Ordinal Regression Model

data. This indicates that its learned features do not progress from poor to good in the manner analogous to actual deterioration of roads. The ORD model, on the other hand, was able to maintain the ordered pattern of t-SNE clusters in both the training and test set feature embeddings, once again displaying its value in modeling ordinal road quality data despite slightly lower raw classification accuracy.

6. Closing Remarks

Our project was ultimately successful in achieving classification accuracy comparable to prior work on applying satellite imagery to road quality evaluation. Futhermore, we have achieved these results on the specific task of assessing a fully paved road network where less reliance can be placed on surrounding urbanization cues. However, more work remains to be done before our model could be deployed in an active role for allocating infrastructure repair resources, as the model was only trained on a 3-category task and still exhibits higher-than-acceptable levels of error for a commercial context.

6.1. Future Work

The primary bottleneck in our project was the quality of the data. In particular, the low resolution of the Google Static Maps imagery often made it difficult to discern the presence and type of road deformities. Higher resolution aerial imagery is available commercially through companies such as Maxar Technologies, but usually this data is not free. Additional funding would be needed to explore the use of high-grade satellite photos.

Dashcam footage could also provide an alternative to the free satellite imagery utilized in our project. Images taken from dashcams have the advantage of being closer to the ground and could capture road features that are difficult to see from a satellite view. Such dashcam footage could also be captured in parallel with mounted sensors on a vehicle driving down a road [10], thus providing multiple synchronized data points for assessing road quality. While firms such as Nexar collect this kind of data, it again comes at a price. Like all good things, high-quality road imagery rarely comes for free.

7. Acknowledgements

We thank Saima Masud for her expertise in the area of PASER evaluation and for providing road condition ratings for 150 randomly selected test set images.

7.1. Contributions

N.A., D.K., and T.M. gathered the ratings data and images. N.A. and D.K. hand-labeled 2,700 images for training

and validation. N.A. created the data augmentation and training pipeline. D.K. implemented the t-SNE clustering visualization pipeline. N.A. and T.M. wrote the paper. T.M. created the project poster.

The following GitHub repositories were referenced for this project:

Image Classification in PyTorch:
https://github.com/spmallick/learnopencv/tree/master/Image-Classification-in-PyTorch

Ordinal Regression:
https://gist.github.com/MathiasGruber/-ef706cc4ede23b239024fec818b201d4

References

[1] E. Brewer, J. Lin, P. Kemper, J. Hennin, and D. Runfola. Predicting road quality using high resolution satellite imagery: A transfer learning approach. *PLOS ONE*, 16(7):e0253370, July 2021.

[2] G. Cadamuro, A. Muhebwa, and J. Taneja. Assigning a grade: Accurate measurement of road quality using satellite imagery. *CoRR*, abs/1812.01699, 2018.

[3] J. Cheng, Z. Wang, and G. Pollastri. A neural network approach to ordinal regression. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. IEEE, June 2008.

[4] D. C. Ciresan, U. Meier, L. M. Gambardella, and J. Schmidhuber. Deep big simple neural nets excel on handwritten digit recognition. *CoRR*, abs/1003.0358, 2010.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[7] J. McBride and A. Siripurapu. The state of u.s. infrastructure. *Council on Foreign Relations*.

[8] V. Mnih. *Machine Learning for Aerial Image Labeling*. PhD thesis, CAN, 2013. AAINR96184.

[9] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[10] G. D. D. Silva, R. S. Perera, C. Keppitiyagama, K. D. Zoysa, N. M. Laxaman, and K. Thilakarathna. Automated pothole detection using wireless sensor motes. 2008.

[11] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2013.

[12] Southeast Michigan Council of Governments. Accessed May 6, 2022.

[13] L. van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[14] D. Walker and L. Entine. *Paser Asphalt Roads Manual*. Transportation Information Center, University of Wisconsin–Madison, 2002.