

COMPARING CORRELATION BETWEEN IMAGE RESOLUTION AND MASK R-CNN MODEL TRAINING TO PREDICT ROAD STRESS INSTANCES.

Kirill Rogovoy, College of Engineering, California State Polytechnic University Pomona, 3801 West Temple Avenue, Pomona, CA, 91768, 781-315-5156, ksrogovoy@cpp.edu

Khac Le, College of Engineering, California State Polytechnic University Pomona, 3801 West Temple Avenue, Pomona, CA, 91768, 626-861-2693, kll@cpp.edu

Patrick Tran, College of Engineering, California State Polytechnic University Pomona, 3801 West Temple Avenue, Pomona, CA, 91768, 714-473-2871, phtran@cpp.edu

Daniel Underwood, College of Engineering, California State Polytechnic University Pomona, 3801 West Temple Avenue, Pomona, CA, 91768, 978-503-4875, dunderwood@cpp.edu

Nancy Lu, College of Engineering, California State Polytechnic University Pomona, 3801 West Temple Avenue, Pomona, CA, 91768, 626-726-7274, nlu@cpp.edu

Wen Cheng, College of Engineering, California State Polytechnic University Pomona, 3801 West Temple Avenue, Pomona, CA, 91768, 909-869-2957, wcheng@cpp.edu

ABSTRACT

Automated detection allows for efficient recognition of road stress. Automated detection of road stress in this study - will be done using a Mask Regional-Convolutional Neural Network (Mask R-CNN) backbone model to train a road stress detection algorithm. The same Mask R-CNN model will be used to predict road stress instances on public roads in three different image resolutions. The image resolutions which are to be compared are 720p, 1080p, and 4k. The authors seek to determine which of the three image resolutions would serve as the most accurate basis for road stress prediction using a Mask R-CNN backbone model.

Keywords: Image Resolution, Mask R-CNN, Road Stress Detection, Neural Networks.

INTRODUCTION

Road surface conditions have a great impact on motorist safety and convenience. Excessive presence of road stress creates hazardous driving conditions, imposes excessive wear upon automobiles, and creates discomfort for drivers. If not addressed in the early stages of development, road stress will continue to progress and consequently present more problems for a motorist, and higher repair costs for a municipality (Zhang & Guo, 2022). That is why it is crucial for municipalities to closely monitor their streets for signs of road stress, and use retained information regarding occurring instances to make necessary repairs. Necessary repairs are based on the type of road stress present, whether it be alligator, transverse or longitudinal cracking (Wynand et al., 2011). That is why these three types of road stress will be examined in this research. For this purpose, images of roads with occurring road stress will be collected from the public streets of the City of Pomona, California. With the development of software such as Google Earth, it is now possible to efficiently and economically collect images of occurring road stress without including extensive labor hours spent on the retention of photographic data points. Due to this efficiency, researchers propose to collect data of roads containing road stress occurrences from the Google Street View (GSV) tool in Google Earth. The collection process entails navigating streets through GSV, selecting areas where road stress is prevalent, and recording the image as well as the location of the image. Depending on location, a major advantage of GSV in large metropolitan areas

where streets are photographed at least once per year, is the ability to save a street image in multiple resolutions, such as 1280x720 (720HD), 1920x1080 (1080HD), and 3840x3160 (4k UHD). The higher an image’s resolution, the more time and processing power will be demanded by the Machine Learning algorithm for instance classification. Due to this, researchers set out to compare the effect that image resolution has on the algorithm’s capability to successfully detect and classify road stress. For the purposes of this study, locations were selected with visible signs of road stress and recorded as saved images in three different resolutions as described above. The frequency of street photography also poses its own benefit; as with yearly photographic recordings of street imaging, it is possible to track the development of road stress over time within a municipality. This is important as the more that a road surface deteriorates over time, the more road stress is developed as a result of the deterioration. Thus, it is important to identify road stress in its early stages of development in order to prevent the further spread and higher repair costs associated with it’s spread. Road stress instance recognition was conducted using a subset of Machine Learning called Computer vision, using a Mask R-CNN backbone model. The particular backbone model used for instance recognition is called R50_FPN_3x. This backbone model was chosen due to the simplicity of training and the existing wide range of the model’s application in road stress detection tasks. (Xu et al., 2022).

DATA DESCRIPTION

For this project, all data was collected through Google Earth. In total, 150 images from 50 distinct locations were gathered. Instances of road stress were identified in all the collected images. Locations for the images were selected primarily from the Pomona, California area, and in rare cases from municipalities close to Pomona. The criteria for location selection included - the presence of road stress on black asphalt pavement, a consistent camera angle, and a consistent zoom distance to the road stress - from the Google Street camera point of view. Once the above criteria was met, images from all locations were saved in resolutions of 1280x720 (720p HD), 1920x1080 (1080p HD), and 3840x3160 (4k UHD), and further used in the annotation process. **Table 1** below portrays an example of a single location, saved in three different image resolutions (from left to right - 720p HD, 1080p HD and 4k UHD).

Table 1 - The same location in three image resolutions

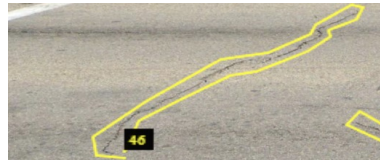
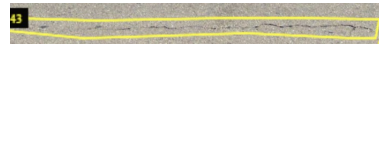
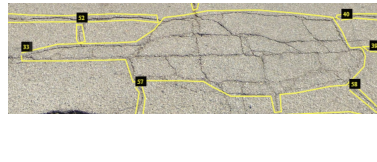
Location - 3, 720p	Location - 3, 1080p	Location - 3, 4k
		

DATA ANNOTATION

The team of researchers delegated tasks for image annotations. Annotations were conducted using the online VGG annotator, strictly using the poly line tool. Researchers were tasked to make consistent identifications of road stress in images of different qualities. Researchers were also tasked with classifying the annotations which they have made into categories of the

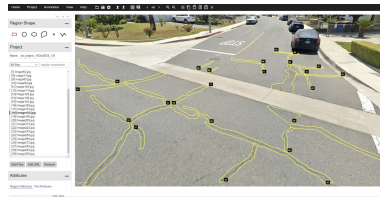
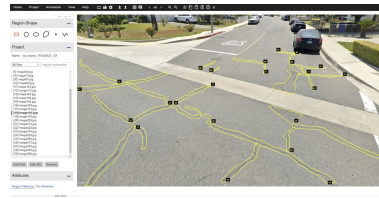
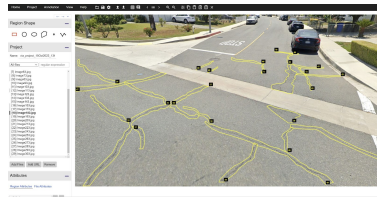
type of road stress encountered within the image. Types of road stress classifications used in the annotation process included alligator, transverse and longitudinal stresses. **Table 2** below portrays the types of annotations made for each of the road stress categories.

Table 2 - Types of occurring road stress and according to annotations.

Transverse Cracking	Longitudinal Cracking	Alligator Cracking
		

In order to further enforce consistency within the results, the same individual would annotate all three images of different resolutions for one location, carefully attempting to make similar annotations to the present road stress through all instances. **Table 3** below portrays screenshots where three images in the qualities of 720p HD, 1080p HD and 4k UHD from a single location, which were annotated in the VGG online annotation tool.

Table 3 - Types of occurring road stress and according annotations.

Location 3, 720p Annotated	Location 3, 1080p Annotated	Location 3, 4k Annotated
		

The number of annotations per image ranged from 6 to over 65, depending on the amount of road stress present within the image.

DATA PROCESSING

Once all 150 images were annotated, researchers set up .csv files with 30 locations and 30 images of the same resolution per .csv file for the first round, and 20 locations with 20 images of the same resolution for the second round. In total, 6 .csv files were created. It is important to note that each .csv file featured images from 20 or 30 locations, but consisted strictly of the same image qualities. For example, 3 .csv files were made for locations 1-30, with each file depicting images of locations 1-30 in different qualities; one .csv file for locations 1-30 with images in 720p HD resolution, another file for locations 1-30 with images in 1080p HD resolution and another .csv file for locations 1-30 with images in 4k UHD resolution. A similar procedure was conducted for locations 31-50. This was done for matters of convenience in organization. The .csv files were then individually uploaded to the VGG annotator tool along with images of corresponding resolutions and locations. From the VGG annotator, .json files were extracted and used for successive training of models.

METHODOLOGY

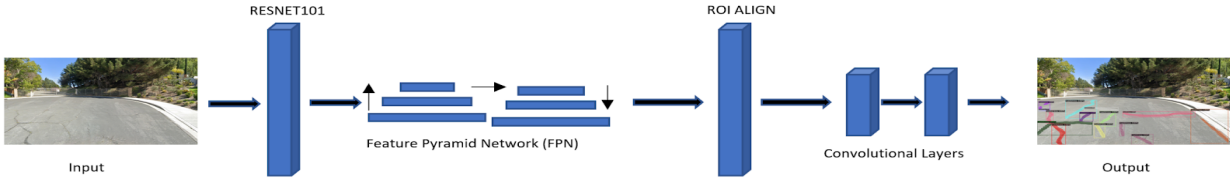
Autonomous Object Detection

Autonomous object detection became more efficient when the concept of Convolutional Neural Networks was developed. Convolutional Neural Networks operate by segmenting an image into multiple pixel matrices, and applying multiple filters, to the said pixel matrices. Each applied filter seeks out a particular characteristic within the moving pixel matrix and records the presence of such a characteristic if identified. When the effort of multiple filters is conjugated, it is possible to yield a comprehensive analysis of the analyzed feature, and eventually create a feature map. The feature map is a result of many dot products between filters and analyzed pixel matrices. It functions to record the locations of where features of interest are identified, by assigning appropriate values to the analyzed image, to rate the importance of a certain area within the feature map. With many feature maps, the algorithm may associate a location on the image with a presence of a desired characteristic, which would mislead the algorithm. In order to prevent this incorrect association, a pooling layer is used. The pooling layer deploys another filter, which notes the presence of desired characteristics found within the map. Pooling layers can operate based on a maximum or average value found within the scope of an applied filter. The maximum pooling layer records the maximum value identified in a region from the convolutional layer's dot product operation and records it on a new map. Meanwhile, the average pooling layer takes the average of surrounding values recorded by the convolutional layer and also maps this information on a new map. The identified higher values are deemed to be of greater importance, so the pooling layer brings the algorithm's attention to regions where higher values are found. Finally, acts the connected layer which combines the work of convolutional and pooling layers to produce a vector, which determines if a feature is present within the analyzed region or not.

Mask R-CNN

Mask Region-Based Convolutional Neural Network (Mask R-CNN) is a model developed for the recognition of objects within an image, and for creating a contour around the said recognized object. Mask R-CNN is the result of progression of instance segmentation technology from Convolutional Neural Networks. In its structure, Mask R-CNN utilizes Feature Pyramid Networks (FPN), Region Proposal Networks (RPN) and Region of Interest Align (ROI Align) to determine the presence of an instance of interest. The FPN extracts important features within an image and maps them onto differently scaled feature maps to track the location of desired instances. Mask R-CNN differentiates significantly from CNN, largely due to having a feature that classifies regions of interest prior to the convolutional filters, hence the R which stands for Region. This region of interest proposal works by analyzing outputs from the pooling layer and creates anchor boxes for areas of high potential interest. The anchor boxes are then split into background and foreground, thus creating bounding boxes and forcing the algorithm to deeper analyze the boxes where a deemed feature is likely to occur. Next, with the assistance of ROI align, the bounding boxes of interest are paired with location (x and y) coordinates to make a more accurate feature map. The approximate operation of the Mask R-CNN algorithm is illustrated in **Figure 1** below.

Figure 1 - Pipeline of the Mask R-CNN algorithm.



Evaluating the Mask R-CNN Model

Once the model has created an output, it is necessary to evaluate the quality of such an output. This is done using loss functions. The general loss function equation is outlined in **Equation 1** below.

$$L=L_{cls}+L_{box}+L_{mask} \quad (1)$$

This equation accounts for losses due to classification (L_{cls}), due to bounding boxes (L_{box}) and loss due to the work of the mask (L_{mask}) - a feature specific to Mask R-CNN. The computation for the mask loss includes a separate formula as outlined in **Equation 2** below.

$$L_{mask} = -\frac{1}{m*m} \sum_{1 \leq i,j \leq m} [y_{ij} \log(\hat{y}_{ij}^k) + (1 - y_{ij}) \log(1 - \hat{y}_{ij}^k)] \quad (2)$$

Training

Training of the road stress recognition model was done following the procedure above and was conducted on an Nvidia GeForce RTX 2080 Super graphics card with 8 GB memory. Certain parameters of the algorithm, specific to this research, which have been used for the training are outlined in **Table 4** below.

Table 4: Hyper-parameter specifications.

Hyper-parameter name	Detectron2's parameter name	Value
Warmup iteration	cfg.SOLVER.WARMUP_ITERS	2000
Base learning rate	cfg.SOLVER.BASE_LR	0.001
Training iteration	cfg.SOLVER.MAX_ITER	100,000
Checkpoint period	cfg.SOLVER.CHECKPOINT_PERIOD	20000
Number of classes	cfg.MODEL.ROI_HEADS.NUM_CLASSES	1
Batch size per image	cfg.MODEL.ROI_HEADS.BATCH_SIZE_PER_IMAGE	512
Anchor sizes	cfg.MODEL.ANCHOR_GENERATOR.SIZES	8,16,32,64,128

Table 4: Hyper-parameter specifications (continued).

Anchor aspect ratio	cfg.MODEL.ANCHOR_GENERATOR.ASPECT_RATIO	0.25,0.5,1.0
Input image width	cfg.INPUT.MIN_SIZE_TRAIN, cfg.INPUT.MIN_SIZE_TEST	600 pixels
Input image's height	cfg.INPUT.MAX_SIZE_TRAIN, cfg.INPUT.MAX_SIZE_TEST	800 pixels

RESULTS

In order to efficiently analyze the results of the study, it is necessary to discuss a phenomenon known as a false positive (FP), true positive (TP) and false negative (FN). A false positive is when a non-present segment of road stress is marked as present, while a true positive is when a present segment of road stress is correctly identified as present. These definitions are used in **Equation 3** and **Equation 4** found below, which are called the precision and recall functions accordingly. These equations are used to quantify the quality of an algorithm's object recognition.

$$precision(p) = \frac{TP}{TP+FP} \quad (3)$$

$$recall(r) = \frac{TP}{TP+FN} \quad (4)$$

Precision quantifies the relationship between the amount of correct positive predictions made from the overall total of made predictions. And, recall quantifies the amount of true positive predictions made from the total amount of predictions that should be true positive. When the precision is mapped on the y-axis and recall is mapped on the x-axis of the same chart, researchers are able to calculate the Average Precision (AP) score - which is simply the area under the resulting curve from the plotting of recall and precision functions. The higher the AP score, the better the performance of the model. The method for calculating the AP score is outlined below in **Equation 5**.

Equation 5:

$$AP = \int_0^1 p(r)dr \quad (5)$$

In **Table 5** provided below, the AP score is provided along with AP50, AP75, AP_s, AP_m and AP_l score values. The AP50 and AP75 values represent the AP scores at a given IoU (which is the area of overlap divided by the area of the union, 50% and 75% regions respectively). While the AP_s, AP_m and AP_l is the AP score for small, medium and large road stress examples respectively (Truong et al., 2021).




Table 5: AP scores for trained models.

Dataset	Model	Bounding Box (Overall)						Segmentation (Overall)					
		AP	AP ₅₀	AP ₇₅	AP _s	AP _m	AP _l	AP	AP ₅₀	AP ₇₅	AP _s	AP _m	AP _l
Location 1-30, 720p	R50_FP-N_3x	37.24	62.03	38.09	20.33	43.58	49.56	18.43	49.52	8.31	7.75	20.87	34.05
Location 1-30, 1080p	R50_FP-N_3x	38.57	64.55	39.99	7.574	38.55	50.55	18.62	49.87	7.45	2.86	17.51	28.07
Location 1-30, 4k	R50_FP-N_3x	36.19	61.60	38.31	0.00	23.84	50.84	11.95	33.09	4.82	0.00	4.49	21.15
Location 31-50, 720p	R50_FP-N_3x	40.27	65.84	42.81	15.61	44.71	61.34	22.15	58.08	10.35	7.32	23.97	41.21
Location 31-50, 1080p	R50_FP-N_3x	46.83	70.41	56.31	1.18	45.05	61.10	24.61	61.03	14.34	2.38	22.54	33.86
Location 31-50, 4k	R50_FP-N_3x	37.69	63.77	39.30	20.05	44.40	50.61	19.03	51.37	8.51	8.19	20.84	36.62

As can be seen in the calculated AP scores from trained models above, for the first round of model runs (Locations 1-30, resolutions 720p, 1080p and 4k), the bounding box AP score was the highest for the 1080p resolution, but not significantly higher than the 720p and 4k resolution. In particular, the 1080p resolution model's bounding box AP score was 3.4% higher than for the 720p model, and 6.2% higher than the 4k resolution. This pattern also applies to the segmentation AP scores, where the 1080p image resolution model received higher AP scores than the models which worked with images of other resolutions. A similar trend is to be noted in the second round of model runs (Locations 31-50, resolutions 720p, 1080p and 4k), where the 1080p model scored the highest in bounding box and segmentation AP scores. In the case of the second set of models, the 1080p resolution attained an AP score that was 14% greater than the 720p model and 19.5% greater than the 4k image resolution.

And once again - the segmentation AP scores were higher in the 1080p image resolution model compared to 720p and 4k. Another piece of information to note - is that the AP50 scores for the models trained were significantly larger than the AP75 scores. This discovery came in line with the researcher's expectations, as the higher IoU threshold in AP75 compared to AP50 means that some true positives, which qualified for the AP50 threshold, are filtered out for the AP75 threshold. Finally, looking over the AP_s , AP_m and AP_l values for all trained models, it is evident from the higher AP_l scores that the model is better at detecting larger instances of road stress. Visual examples of predictions made by the model can be seen in **Table 6** below.

Table 6 - Predictions made by the model for three image resolutions in Location 24.

Resolution 720p Annotated	Resolution 1080p Annotated	Resolution 4k Annotated
		

CONCLUSION

From an analysis of the above information, it is evident that the model which ran predictions on the images of 1080p resolution produced the highest AP scores overall, and thus yielded the best results. Based upon this, the researchers have concluded that the 1080p image resolution provides a superior image format for training a road stress recognition model, on a Mask R-CNN backbone. In this study, the three most commonly encountered image resolutions were studied, but for future studies, researchers would be interested in examining the effect of significantly lower or higher than 1080p image resolution, to determine whether there is a significant variance in the quality of the model's predictions. But, based on the conducted experiment, the 1080p image resolution seems to be the perfect middle ground for the selection of the ultimate image size to train a Mask R-CNN model.

REFERENCES

- Guo, & Zhang, Z. (2022). Road damage detection algorithm for improved YOLOv5. *Scientific Reports*, 12(1), 15523–15523. <https://doi.org/10.1038/s41598-022-19674-8>
- Kirill Rogovoy. (2021). Road Stress Detection Using Mask R-CNN. California State Polytechnic University, Pomona.
- Liu, Yang, G., Huang, Y., & Yin, Y. (2021). SE-Mask R-CNN: An improved Mask R-CNN for apple detection and segmentation. *Journal of Intelligent & Fuzzy Systems*, 41(6), 6715–6725. <https://doi.org/10.3233/JIFS-210597>
- Maeda, Sekimoto, Y., Seto, T., Kashiya, T., & Omata, H. (2018). Road Damage Detection and Classification Using Deep Neural Networks with Smartphone Images: Road damage detection and classification. *Computer-Aided Civil and Infrastructure Engineering*, 33(12), 1127–1141. <https://doi.org/10.1111/mice.12387>
- Sheng Tan. (2021). Recognition of Roadway Sign at Night Using Thermal Imaging and Computer Vision. California State Polytechnic University, Pomona.
- Steyn, Liu, X., Mehta, Y., & You, Z. (2011). Road Pavement and Material Characterization, Modeling, and Maintenance edited by Wynand JvdM Steyn, et al. (Steyn, X. Liu, Y. Mehta, & Z. You, Eds.). American Society of Civil Engineers.
- Truong, L., Cheng, W., Clay, E., Mora, O. E., Singh, M., & Jia, X. (2022). (rep.). *Rotated Mask RCNN Detection for Parking Space Management System*. TRB 101st Annual Meeting and Potential Publication at Transportation Research Record.
- Truong, L. N. H., Clay, E., Mora, O. E., Cheng, W., Kaur, M., & Tan, S. Semi-Supervised Learning and Deep Neural Network on Detection of Roadway Cracking Using Unmanned Aerial System Imagery. *Available at SSRN 3988127*.
- Truong, L. N. H., Mora, O. E., Cheng, W., Tang, H., & Singh, M. (2021). Deep Learning to Detect Road Distress from Unmanned Aerial System Imagery. *Transportation Research Record*, 03611981211004973.
- Xu, X., Zhao, M., Shi, P., Ren, R., He, X., Wei, X., & Yang, H. (2022). Crack detection and comparison study based on faster R-CNN and mask R-CNN. *Sensors*, 22(3), 1215. [doi:https://doi.org/10.3390/s22031215](https://doi.org/10.3390/s22031215)