Evaluating Road Segmentation Performance in Participatory Sensing: Investigation into Alternative Metrics

Jeongho Hyeon[1], Minwoo Jeong[1], Giwon Shin[1], Wei-Chih Chern[2], Vijayan K. Asari[2], and Hongjo Kim[1]

[1]Yonsei University, Republic of Korea
hyeon9404@yonsei.ac.kr, 2016144045@yonsei.ac.kr, giwone1330@yonsei.ac.kr, hongjo@yonsei.ac.kr [2]University of Dayton, OH, United States
chernw1@udayton.edu,vasari1@udayton.edu

Abstract

For road condition assessment, participatory sensing has been proposed in literature utilizing a normal vehicle equipped with a dashboard camera. In such environment, the main technical challenge is not only the recognition performance on target classes such as cracks, but also the preparation of training and test datasets with high quality annotations. This study found that the annotation quality presents a unique problem in the performance test of participatory sensing-based road condition assessment. To address the problem, this study explores the adequacy of most commonly-used evaluation metric, the Intersection over Union (IoU), and suggest alternative metrics for road segmentation models in the context of participatory sensing. Experiments were conducted on the AIM crack dataset collected from urban road environments using dashboard cameras on normal vehicles. This study provides new insights into the importance of considering proper evaluation metrics in participatory sensing-based infrastructure monitoring.

Keywords –
Participatory sensing, road crack segmentation, performance evaluation metric, convolutional neural network, semantic segmentation

1 Introduction

The expenses for the maintenance and repair (M&R) of road networks in South Korea have increased significantly, recording $2 billion in 2015 to $2.9 billion in 2021 [1]. Previous research has shown that proactive M&R of road damages can significantly reduce M&R costs compared with reactive M&R [2]. To facilitate proactive M&R, it is essential to monitor road damages in a timely manner. However, limited budgets and monitoring resources in governmental agencies responsible for managing road infrastructure often make it difficult to identify road surface damages promptly.

A previous study suggested an alternative monitoring method for road surfaces through participatory sensing, which leverages the data collection capabilities of citizens [3,4]. The previous study [3,4] utilized deep CNN to identify surface cracks, demonstrating the applicability of AI-based recognition systems. Most recent studies have used object detection models to find road surface damages [5]. However, to make data-informed decisions in road M&R, damage information in a bounding box format could be insufficient as it cannot provide the length and shape of cracks, which are crucial for selecting proper treatment methods.

Semantic segmentation models are advantageous for road M&R as they can recognize target classes, such as cracks, at the pixel-level. The segmentation results naturally facilitate the quantification of crack ratios in road sections, which can help make M&R decisions based on real-inspection data in a timely manner. However, participatory sensing environments presents a unique challenge in preparation of training and test datasets with accurate annotations due to poor image quality. Poor image quality is attributed to the low specification of dashboard camera image sensors, car motions, external lighting conditions, and weather effects. Therefore, annotated images often have erroneous annotations as shown in the right-hand side of Fig. 1. This problem is not only caused by human error, but also poor image quality with excessive noise as shown in Fig. 2. With such image quality, annotators are hard to differentiate cracks from road surfaces. Additionally, as shown in Fig. 1, the large field of view captured by a dashboard camera leads to objects in the distance appearing smaller and narrower, making the annotation task challenging.

When it comes to making decisions for proactive M&R of road networks, critical information is the ratio and length of cracks in a road unit. Although there is a widely used road condition indicator such as the international roughness index (IRI), proactive M&R activities do not need such level of details. Therefore, in the context of participatory sensing, the required level of details for road conditions includes the shape of cracks, not the width of them. Considering this, the current evaluation metric, the IoU, could mislead the interpretation of experimental results of semantic segmentation models, as it checks how accurately the model predicts crack regions in pixels. Fig. 1 illustrates this problem. From a road manager's perspective, the prediction result is acceptable as it can provide the current status of road cracks, but the IoU score is low recording 0.28. Few pixel differences in location, thickness, and length between the predicted results and ground truth have little effect on road M&R judgment.

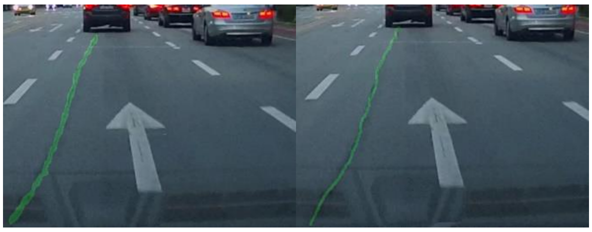


Fig. 1. An example of the segmentation results (left) and the ground truth (right). The IoU Score is 0.28.

To address this issue, this study investigates alternative performance evaluation metrics for participatory-sensing based road M&R decision making. The first evaluation method is composed of a morphological operation and a buffer method to yield performance scores for evaluation metrics such as 'completeness', 'correctness', and 'quality'. The buffer method is a simple matching procedure in which any portion of the prediction pixel within a defined pixel distance from the ground truth (GT) pixels is considered as a correct match. The second method, the keypoint matching method, evaluates the distance between the prediction and ground truth keypoints. The proposed evaluation metrics can be useful for road monitoring in the context of participatory sensing as the goal of crack recognition is identifying the ratio and length of cracks in certain road units to facilitate proactive M&R for road networks.

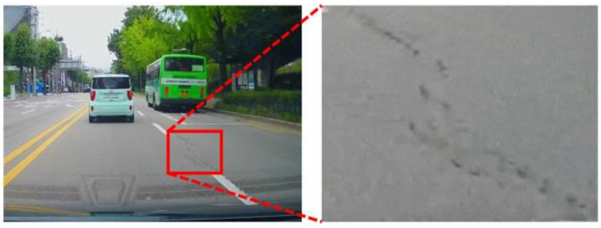


Fig. 2. Poor image quality due to harsh imaging conditions using a dashboard camera in a moving car.

2 Evaluation metrics for participatory sensing-based road monitoring

New evaluation metrics for participatory sensing-based road monitoring are designed to achieve the following goals:

- The performance of road crack segmentation is evaluated by the crack shape and length.
- The segmentation results should be meaningful to make decisions for proactive M&R of road networks.

To achieve the goals, this study designs and examines two evaluation metrics.

2.1 Shape-based evaluation of crack segmentation

The first evaluation method is named as 'shape-based evaluation of crack segmentation', as it focuses on the integrity of crack shapes. That is, the metric evaluates whether a segmentation model correctly recognize the crack shapes, with the tolerance on the difference between the pixel thickness of prediction results and the ground truth. For example, if a prediction result is a straight crack line with the pixel width of 5 and length of 100, and the ground truth is also a straight crack line with the pixel width of 1 and length of 100, then the metric should give 100% score.

To realize the above idea, this study adopts CCQ metric which utilizes the concept of 'completeness', 'correctness', 'quality', and the buffer [6]:

'Completeness' represents the proportion of the predicted cracks that lies within the buffer around the ground truth cracks. It is expressed as a percentage and defined by equation (1).

'Correctness' refers to the proportion of the predicted cracks that lie within the buffer of the ground truth network, as quantified by equation (2).

'Quality' of the result, which takes into account both completeness and correctness, is a measure of how well the final outcome has been achieved. It is quantified by equation (3) and it is equivalent to IoU.

"F1 score" is the harmonic mean of completeness and correctness, and it is defined by equation (4).

$$Completeness = \frac{TP_{buffer}}{TP_{buffer} + FN_{buffer}} \tag{1}$$

$$Correctness = \frac{TP_{buffer}}{TP_{buffer} + FP_{buffer}} \tag{2}$$

$$Quality = \frac{TP_{buffer}}{TP_{buffer} + FP_{buffer} + FN_{buffer}} \tag{3}$$

$$F1score = \frac{2 * Completness * Correctness}{Completness + Correctness} \quad (4)$$

The term "True Positive" (TP) refers to the correct identification of a crack segment by the model. A "False Positive" (FP) refers to a pixel that the model falsely identified as a crack, while a "False Negative" (FN) refers to a real crack pixel that the model failed to identify as such.

Fig. 3 illustrates the buffer concept. Based on GT, it is defined as TP if the prediction is included in the buffer width, and FP if not. If GT is not included in the buffer width based on the prediction, it is defined as FN. In this study, the buffer width was defined as 5 pixels. In other words, if the predicted result is included within GT and 5 pixels, it is determined that the prediction is successful.

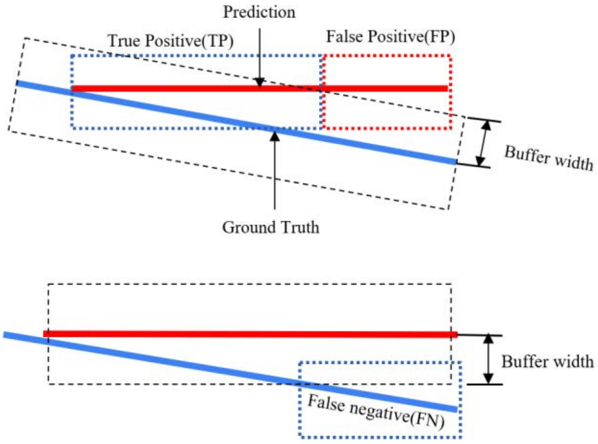


Fig. 3. Illustration of TP, FP, and FN of cracks using CCQ metric with the buffer

Skeletonization is also applied to predicted and ground truth masks of cracks to examine the segmentation performance focusing only on the crack shape. Skeletonization is a morphological operation in image processing that aims to produce the skeletal structure or the thinned version of a binary image, preserving only the essential structures of the image. The goal of using skeletonization is to obtain a simplified representation of road cracks while retaining their shape and topology.

2.2 Keypoint-based evaluation metric

The second evaluation metric examines the keypoint matching between predictions and ground truth, to mitigate the sensitivity to pixel thickness. To this end, this study employed the GFTT (Good Features to Track) concept, introduced by Shi and Tomasi in 1994. Their objective was to determine which features are appropriate for tracking in a feature-based vision system, as the identification and tracking of good features is crucial for its operation (Shi and Tomasi 1994). By adopting this features, the shape-based evaluation is possible for crack segmentation results. The GFTT detector was originally designed to extract features such as corner points, similar to the key point extraction method of Harris and Stephens [5]. The coordinates of the keypoints were obtained from the segmentation results and ground truth images. If a keypoint of the segmentation results is within 5 pixels of a keypoint of the ground truth, it was considered a True Positive ($TP_{prediction}$). If a keypoint of the ground truth is within 5 pixels of a keypoint of the segmentation results, it was considered a True Positive (TP_{GT}). Prediction, GT, and Keypoint scores are defined by equation (5), (6), (7).

$$Prediction\ score = \frac{TP_{prediction}}{\#\ of\ keypoints\ in\ Prediction} \quad (5)$$

$$GT\ score = \frac{TP_{GT}}{\#\ of\ keypoints\ in\ GT} \quad (6)$$

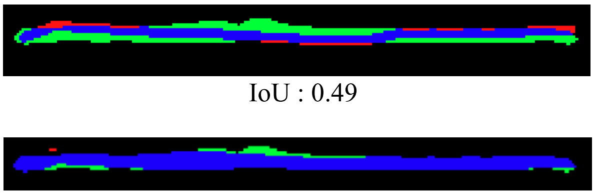
$$Keypoint\ score = \frac{2 \times Prediction\ score \times GT\ score}{Prediction\ score + GT\ score} \quad (7)$$

3 Experiment

The experiments were conducted on the AIM crack dataset [3], collected by multiple normal vehicles equipped with dashboard cameras. The dataset consisted of 327, 100, and 100 images for training, validation, and test, respectively. The dataset provides polygon annotations for road cracks. DeepLabv3+ [8] and FPN [9] were used as a segmentation model. In the implementation, DeepLabv3+ has been trained for 50 epochs with the encoder part of mit_b5, which was SegFormer [10] pre-trained on ImageNet. The mini-batch had 6 images, the learning rate of 0.0001 before 25 epochs and 0.00001 after, and the optimizer of Adam were set to train the model. For comparison, FPN was trained for 50 epochs with the encoder part of Efficientnet_b5[11], which was pre-trained on ImageNet. The mini-batch had 16 images, the learning rate of 0.0001 before 25 epochs and 0.00001 after, and the optimizer of Adam were set to train the model.

3.1 CCQ metric experiment

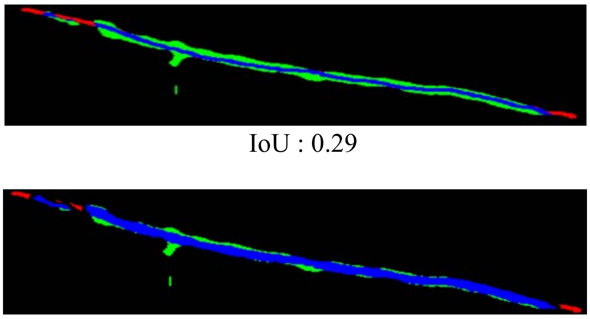
To count TP, FP, and FN, this study produced the crack masks on the test images (see examples in Figs. 1, 8 and Table 1). As shown in Fig 4, the location and shape of the cracks were accurately predicted. However, the presence of FP suggests that the model's prediction of crack thickness is wider than the ground truth. This result yields a low IoU score of 0.49. In Fig 5, for the same reason, IoU score was only 0.29. As the correct prediction of crack thickness is less critical in making proactive M&R decisions for road networks, the shape-based evaluation using the CCQ score and the skeletonization was conducted, as shown in Fig 6, 7. The CCQ score provides insight into the model's ability to produce results that are similar to human cognitive evaluation in the context of participatory sensing, by using the buffer method.



IoU : 0.49

Completeness:0.9, Correctness:0.89, Quality:0.81

Fig. 4. Segmentation mask visualization of IoU (top) and the proposed metrics (bottom) using DeepLabV3+. TP is blue, FP is green, FN is red.

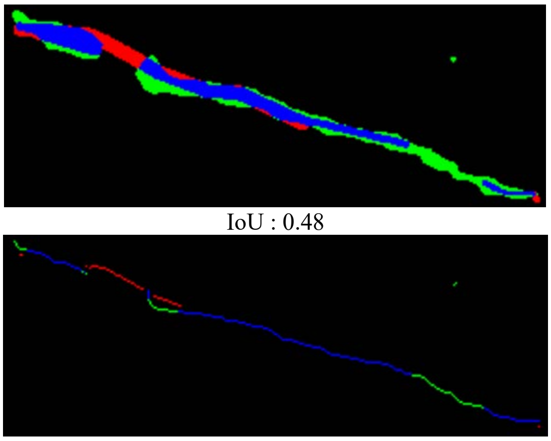


IoU : 0.29

Completeness:0.95, Correctness:0.62, Quality:0.60

Fig. 5. Segmentation masks produced by FPN, with the IoU (top) and proposed metrics (bottom). TP is blue, FP is green, FN is red.

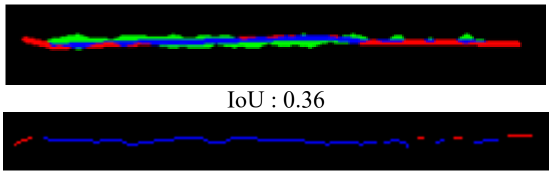
To further reduce errors due to the crack thickness difference, this study performed skeletonization on both prediction and ground truth masks, and measured CCQ again. Skeletonization is a critical preprocessing step for crack segmentation models, as it can mitigate errors arising from differences in prediction and ground truth crack thickness. For instance, if a crack is labeled with a thickness of one pixel in the ground truth, an accurate prediction of the crack's location with a thickness of four pixels would yield an IoU score of only 0.25. By skeletonizing both prediction and ground truth mask, the thickness is reduced to one pixel(Fig 8), thereby enabling better correspondence with the ground truth and improving the accuracy of the segmentation model. Skeletonization was performed using the Python package called scikit-image. Using this package, the skeletonization algorithm iteratively removes peripheral pixels of the image data, while preserving the connectivity of the objects of interest. This operation is performed iteratively until the desired level of a one-pixel thick skeletonization is achieved.



IoU : 0.48

Completeness:0.8, Correctness:0.62, Quality:0.54

Fig. 6. Segmentation mask visualization of IoU (top) and the CCQ metric after skeletonization (bottom) using DeepLabV3+. TP is blue, FP is green, FN is red.



IoU : 0.36

Completeness:0.87, Correctness:0.91, Quality:0.8

Fig. 7. Segmentation masks produced by FPN, with the IoU (top) and the CCQ metric after skeletonization (bottom). TP is blue, FP is green, FN is red.

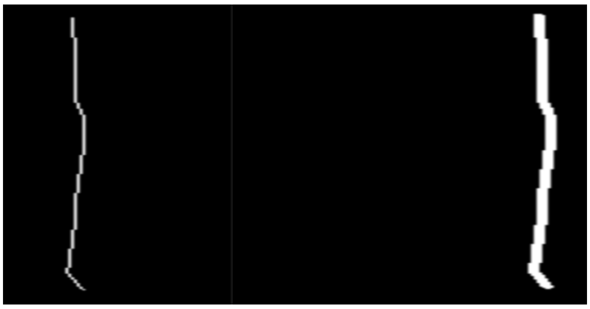
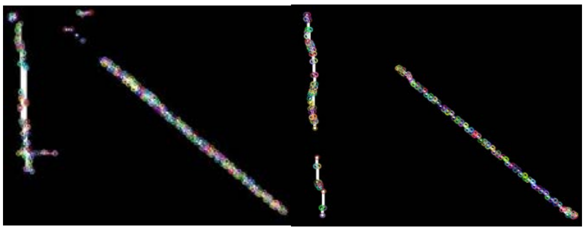


Fig. 8. Example of Skeletonization

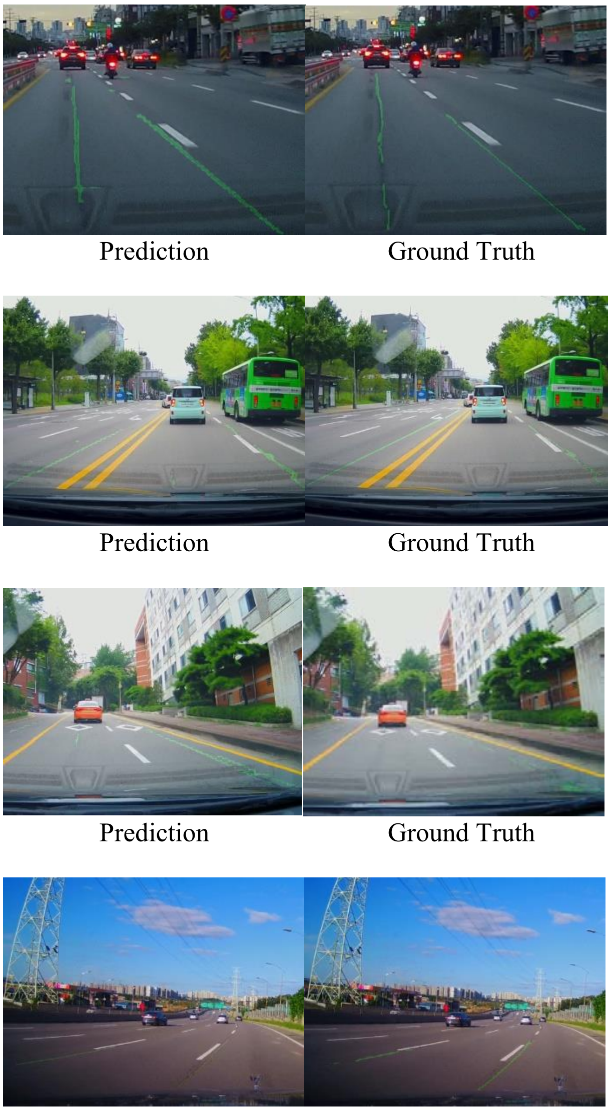
3.2 Keypoint metric experiment

Key point extraction was performed on both the prediction and ground truth masks, and the extracted key points can be seen in Fig. 9. Key points were generated throughout the crack, and matching was performed based on the distance between the key points of the prediction and ground truth. Key points within 5 pixels were assumed to be matched, so the score can be adjusted to compensate for the position and thickness difference of the cracks in the prediction and ground truth mask.



Prediction Ground Truth

Fig. 9. An example of keypoint matching results. (IoU = 0.33, keypoint matching score = 0.6)



Prediction Ground Truth

Prediction Ground Truth

Prediction Ground Truth

Prediction Ground Truth

Fig. 10. Visualization of the predicted masks versus the ground truth masks. Despite correct predictions in crack shapes, the IoU score is low as shown in Table 1.

Table 1. Model performance evaluation in each metric for Fig. 10

Row	IoU	Comp.	Corr.	Qual.	Keypoint
First	0.33	0.92	0.73	0.81	0.6
Second	0.35	0.84	0.78	0.81	0.79
Third	0.24	0.69	0.69	0.69	0.69
Fourth	0.32	0.88	0.89	0.88	0.68

Table 2. Model evaluation

	S*	IoU	Comp.	Corr.	Qual.	Keypoint
Score (Deep LabV3+)		0.22	0.71	0.54	0.43	0.44
	✓	0.06	0.73	0.55	0.44	0.45
	S*	IoU	Comp.	Corr.	Qual.	Keypoint
Score (FPN)		0.24	0.74	0.67	0.53	0.44
	✓	0.07	0.69	0.73	0.53	0.45

S*: Skeletonization

4 Limitation & Future study

In this study, we present a metric that is capable of assessing the predictive capabilities of a road crack segmentation model in participatory sensing environments. The effectiveness of this metric was demonstrated through the experiments. However, the prediction results for fatigue cracks, as shown in Fig. 11, is relatively difficult to handle using the proposed metric. Although the prediction result seems plausible to the naked eye, evaluating the model's performance using the CCQ metric with and without skeletonization was not successful in the case of fatigue cracks, as shown in Fig. 13 and Table 3.

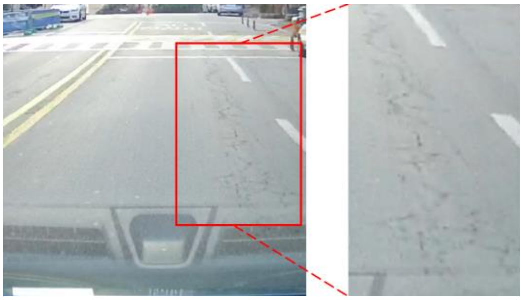
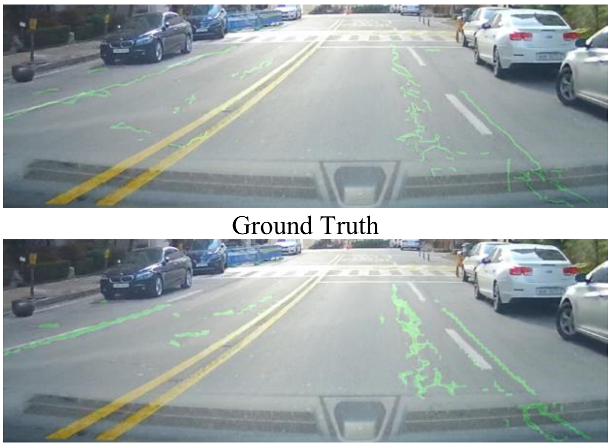


Fig. 11. Example of fatigue crack

Future research is needed to develop a suitable approach for evaluating the model's performance in fatigue cracks. The proposed metric can be applied exclusively to non-fatigue cracks by using object detection models which identify fatigue cracks. If suitable evaluation metrics are applied for each type of crack, it will be possible to more accurately evaluate the model's predictive ability.



Ground Truth

Prediction

Fig. 12. The ground truth and the prediction results were overlaid on the raw images.

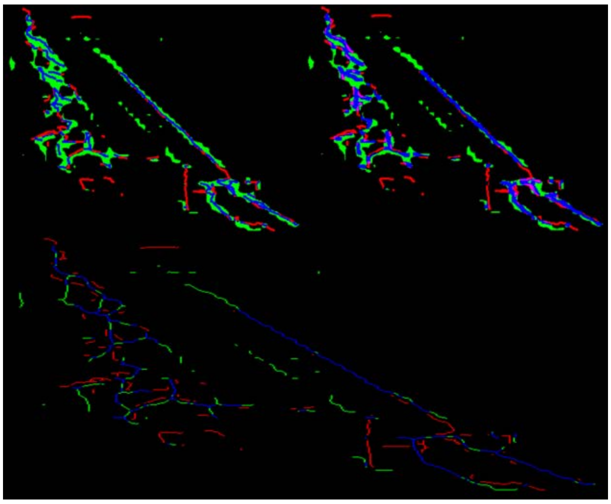


Fig. 13. IoU (Top Left), CCQ (Top Right), CCQ after Skeletonization (Bottom). TP is blue, FP is green, FN is red.

Table 3. Model performance evaluation for fatigue cracks shown in Fig. 13

	S*	IoU	Comp.	Corr.	Qual.
Score		0.26	0.62	0.53	0.4
	✓	0.05	0.49	0.56	0.35

5 Conclusion

This study investigated alternative segmentation performance evaluation metrics for road crack segmentation in the context of participatory sensing. As inaccurate annotations are inevitable due to poor image quality of a dashboard camera, the proposed evaluation metric allow users to examine the segmentation performance with a criterion that the predicted masks are useful for decision-making in proactive M&R for road networks. The experimental results imply the use of IoU score is disadvantageous in participatory sensing, as it is highly sensitive to the width of cracks rather than the predicted crack shape which can provide the crack ratio and length information. Rather than IoU, the CCQ score or keypoint based assessment is more preferable. This study highlights the importance of using appropriate performance evaluation metric to assess road crack segmentation models focusing on crack shapes, thereby facilitating the accurate performance evaluation for participatory sensing-based road monitoring results.

Acknowledgements

This research was conducted with the support of the "2022 Yonsei University Future-Leading Research Initiative (No. 2022-22-0102)" and the "National R&D Project for Smart Construction Technology (RS-2020-KA156488)" funded by the Korea Agency for Infrastructure Technology Advancement under the Ministry of Land, Infrastructure and Transport, and managed by the Korea Expressway Corporation. This work was also supported by K-water Grant funded by the Korean Government(Innovative talent nurturing project in the digital water industry). Any opinions and findings in this paper are those of the authors and do not necessarily represent the funding agencies listed above.

References

[1] Korea Institute of Civil Engineering and Building Technology. Road Statistics and Maintenance Information System. On-line: http://www.rsis.kr/maintenance_summary.htm., Accessed: 30/01/2023

[2] Federal Highway Administration. Pavement Performance Measures and Forecasting and The Effects of Maintenance and Rehabilitation Strategy on Treatment Effectiveness (Revised). U.S. Department of Transportation: Research, Development, and Technology Turner-Fairbank Highway Research Center, 6300 Georgetown PikeMcLean, VA 22101-2296, 2017

[3] Bang, S., et al., Encoder–decoder network for pixel-level road crack detection in black-box images.

Computer-Aided Civil and Infrastructure Engineering, 34(8): p. 713-727, 2019.

[4] Somin Park, Seongdeok Bang, Hongjo Kim, and Hyoungkwan Kim., Patch-based Crack Detection in Black Box Images using Convolutional Neural Networks. *Journal of Computing in Civil Engineering*, 33(3), 04019017, 2019.

[5] Cao, M.-T., et al., Survey on performance of deep learning models for detecting road damages using multiple dashcam image resources. *Advanced Engineering Informatics* 46: 101182, 2020.

[6] Wiedemann, Christian, et al. Empirical evaluation of automatically extracted road axes. *Empirical evaluation techniques in computer vision,* 12: 172-187, 1998.

[7] Zhu, Zhenhua, and Khashayar Davari. Comparison of local visual feature detectors and descriptors for the registration of 3D building scenes. *Journal of Computing in Civil Engineering* 29.5: 04014071, 2015.

[8] Chen, L.-C., et al., Rethinking atrous convolution for semantic image segmentation. *arXiv preprint* arXiv:1706.05587, 2017.

[9] Lin, Tsung-Yi, et al. Feature pyramid networks for object detection. *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

[10] Xie, E., et al., SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 2021. 34: p. 12077-12090.

[11] Tan, Mingxing, and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *International conference on machine learning*. PMLR, 2019.