Crack Detection and Localization in Stone Floor Tiles using Vision Transformer approach

Luqman Ali [a,e,f], Hamad Aljassmi [b,f], Medha Mohan Ambali Parambil [c,e], Muhammed Swavaf [a,e], Mohammed AlAmeri [d], Fady Alnajjar [a,f,*]

[a]Department of Computer Science and Software Engineering, College of IT, United Arab Emirates University (UAEU), Al Ain 15551, UAE
[b]Department of Civil Engineering, College of Engineering, UAEU, Al Ain 15551, UAE
[c]Department of Information Systems and Security, College of IT, UAEU, Al Ain 15551, UAE
[d]Department of Electrical Engineering, College of Engineering, UAEU, Al Ain 15551, UAE
[e]AI and Robotics Lab (Air-Lab), UAEU, Al Ain 15551, United Arab Emirates
[f]Emirates Center for Mobility Research, UAEU, Al Ain 15551, United Arab Emirates
E-mail: 201990024@uaeu.ac.ae; h.aljasmi@uaeu.ac.ae; medhamohanap@uaeu.ac.ae; swavafup@uaeu.ac.ae; 201731009@uaeu.ac.ae; fady.alnajjar@uaeu.ac.ae

Abstract

Cracks are the initial indicators of the deterioration of any civil infrastructure. Structures are typically monitored manually by inspectors, which is time-consuming, laborious, costly, and easily prone to human error. To address these limitations this paper aims to present a vision transformer-based stone floor tiles crack detection and localization approach. The proposed model is trained on a custom dataset acquired from various stone tiles under various illumination conditions in the United Arab Emirates. The dataset consists of 5800 images having a resolution of 224×224 pixels. To assess the effectiveness of the proposed model, various evaluation metrics such as testing accuracy, precision, recall, F1 score, and computational time are employed to analyze its performance. The input patch size of the Vision Transformer (ViT) model is varied to investigate its effect on the performance of the model. The experimental results show that input patch size has a significant on the performance of the models. The ViT model trained on the lowest patch size of 14×14 pixels achieved the highest testing accuracy, precision, recall, and F1 score of 0.8612, 0.8840, 0.8304, and 0.8564 respectively. The inference time of the ViT model for a single patch is 0.092 sec. The crack localization is performed by combining the proposed trained ViT model with the sliding window approach. The model performed well in detecting and locating cracks in stone floor tiles, indicating its potential for practical use.

Keywords

Structural Health Monitoring; Crack Detection; Vision Transformer; Sliding Window Approach; Stone floor; automatic inspection.

1 Introduction

In structural health monitoring, it is important to detect and monitor surface cracks early for long-term maintenance and failure prediction. The structure's condition information can be collected manually by subjective human experts by visually inspecting and evaluating the structure or automatically by using various vision-based approaches. Manual Inspection techniques are laborious, time-consuming, inspector dependent, and easily vulnerable to the perspicacity of the inspector. In addition, numerous studies have demonstrated the inherent variability and inconsistency of visual inspection results [1], [2]. Inadequate inspection and condition assessment can result in various accidents, such as the Minneapolis Interstate 35W Bridge Collapse, which resulted in 13 fatalities and 145 injuries [3]. Another example is the November 28, 1999, incident involving a freight train in Japan's Rebunhama Tunnel, which occurred because shear cracks in the structure were not correctly detected [4]. Automatic inspection techniques provide an efficient solution by reducing subjectivity and providing a substitute for the human eye to circumvent the issues associated with manual inspection. The automatic vision-based crack detection approaches can be divided into traditional image processing, Machine Learning (ML), and Deep Learning (DL)-based approaches. The conventional image processing approaches include various edge detection [5], thresholding [6], and filtering approaches [7]–[10] however these approaches cannot show resilience to image illumination and require manual human efforts. Machine learning approaches have the capability to overcome the limitation associated with conventional crack detection approaches. Machine learning approaches consist of feature extraction i.e., extracting useful features [11]–[13] from images, and classification i.e., classifying the feature into crack and non-cracks using classifiers [14]–[16]. The limitation of the machine learning approaches is the manual selection of feature extraction techniques which is not only challenging but sometimes the extracted features did not represent the actual cracks. These limitations can be addressed by using DL algorithms that are capable of automatic feature extraction and classification. DL algorithms are particularly effective in detecting features in images because they can automatically learn the feature representations from the data itself. Various works [17]–[21] have been presented in the literature using various DL based Convolutional Neural Network (CNN) models for crack detection in civil structures. However, these approaches suffer from localized receptive field problems in which the feature are not extracted in a global context. Vision transformers are a relatively new type of neural network that have shown promising results in image classification tasks, including crack detection. Several research studies have proposed using vision transformers to overcome the limitations of traditional methods in detecting and segmenting cracks [22]–[24]. Vision transformers are particularly effective due to their ability to capture long-range dependencies in images using self-attention mechanisms [25]. Therefore, this paper aims to present a vision transformer-based crack detection in stone floor tiles. The contribution of the proposed work is as follows:

1. A custom dataset of 5800 stone floor tiles with crack and non-crack images having a resolution of 224*224 is created.
2. A first ViT-based framework is proposed for crack detection in stone floor tiles.
3. The performance of the proposed ViT is compared based on various input patch sizes to select an optimum input patch size.
4. Based on the results, a detailed discussion is conducted to provide a reference to a researcher working in the same field.

The remainder of the paper is organized as follows. Section 2 explains the system overview. The experimental results are discussed in section 3 followed by a discussion and conclusion in the last section.

2 System Overview

The proposed ViT-based stone floor tile crack detection system is composed of three main phases, as shown in Figure 1. A dataset is constructed in the first phase, which is subsequently provided to ViT transformer for training. The final section phase of the proposed method involves testing the system after training and performing crack localization using a sliding window approach.

2.1 Dataset Creation

In the proposed work, the data is collected from a variety of stone floor tiles used in the United Arab Emirates. A mobile phone camera with a resolution of 4000x3000 pixels is used for the acquisition of images. The acquired images are divided into small patches of 224 x 224 pixels. In order to separate the acquired data into cracks and non-cracks, manual labeling is performed. A total of 5800 patches are distributed 50/50 between cracks and non-cracks. Figure 2 below shows sample images of the acquired dataset and patches. A data split of 60:20:20 is kept between the patches for training, validation, and testing of the proposed ViT model. The 60:20:20 split provides a balance between using enough data to train the model effectively, tune the hyperparameters of the model, while also ensuring that the model's performance is reliable on new data.

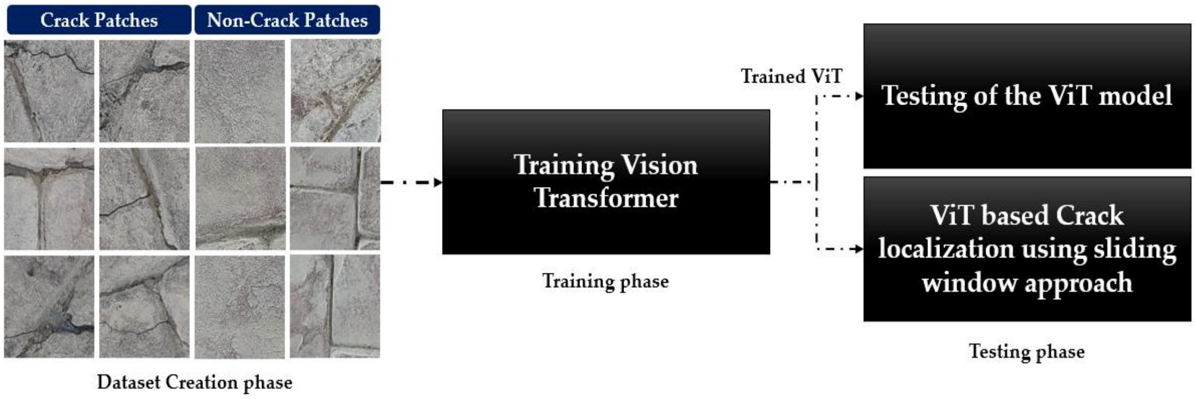

Figure 1. Overview of the proposed system

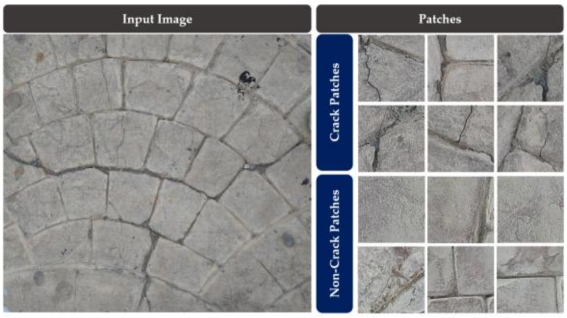

Figure 2. Sample of the acquired image and input patches (Crack and Non-Crack)

2.2 Training of the ViT Model

The data created in the previous phase is given to the ViT model for training purposes. The vision transformer model is first developed by [26] and can overcome the shortcoming of the DL approaches by considering the input images as a series of patches. The schematic diagram of the vision transformer is shown in Figure 03. ViT architecture consists of an embedding layer, an encoder, and a final head classifier. The embedding layers divide an image X from the training set into patches without overlap where each patch is considered a unique token. In the encoder part of the vision transformer, multi-head self-attention (MHSA) is used to extract and integrates information globally across multiple regions of the images whereas traditional CNN uses filters with a local receptive field. This acquired information is encoded and fed into a multilayer perceptron classifier for classification purposes. Interested readers are referred to [26] for more information about the working of the vision transformer. In the proposed work, the ViT model is trained on various patch sizes i.e., 14*14, 21*21, 28*28, and 56*56 to evaluate its effect on the crack detection performance of the ViT model. The hyperparameters are tuned based on trial and error basis. During the hyperparameters tuning stage of the model the learning rate, transformer layers, batch size, and the number of epochs is set to 0.001, 16, 16, and 40 respectively.

2.3 Testing of ViT classifier

The trained ViT classifier is tested with a new set of data that is not used in the training and validation phase of the model. The trained ViT model is also integrated with the sliding window approach to localize the crack region in the images. In the sliding window approach, a window size of 224*224 equal to the input patch size is considered. The single window patch is given to the trained ViT model to identify the crack and non-crack patches. The sliding window moves over 224 pixels horizontally and vertically until the whole image is covered. A red bounding box is drawn around the patch classified as a crack by the ViT model as depicted in Figure 4. Interested readers are referred to [27] for more information about the working of the sliding window approach.

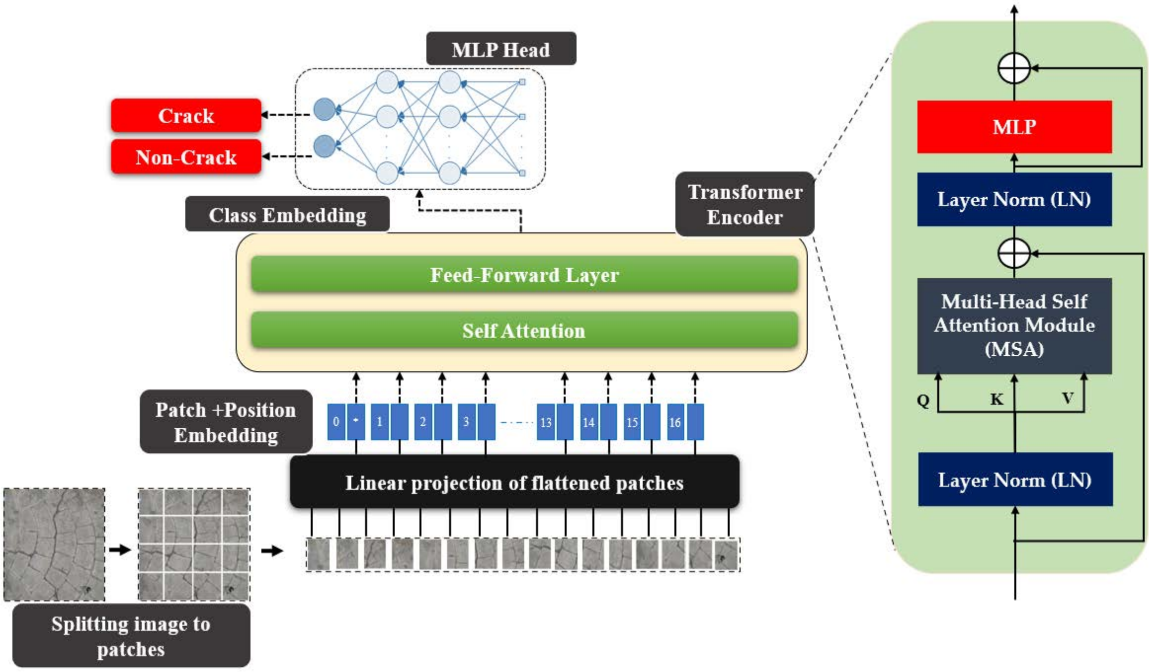

Figure 3. Schematic diagram of vision transformer.

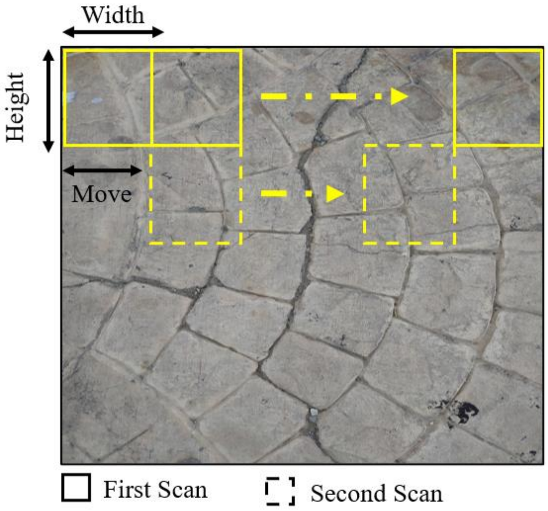

Figure 4. Representation of sliding window approach.

2.4 Evaluation Metrics

To evaluate the performance of the proposed crack detection model, several standard evaluation metrics, including accuracy, precision, recall, and F1 score (depicted in Equation 1, 2, 3, and 4 respectively) are used. Accuracy measures the proportion of correctly classified samples (both cracked and non-cracked) among all samples in the dataset.

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN}) \tag{1}$$

Where TP, TN, FP, and FN represent the true positives, true negatives, false positives, and false negatives respectively. Precision measures the proportion of correctly classified cracked samples among all samples classified as cracked.

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) \tag{2}$$

Recall measures the proportion of correctly classified cracked samples among all true positive samples in the dataset.

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN}) \tag{3}$$

F1 score is the harmonic mean of precision and recall, and is often used to balance the trade-off between the two metrics.

$$\text{F1 score} = 2 * ((\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})) \tag{4}$$

3 Experimental Results

The proposed ViT model is trained on Alienware Arura R8 core i9-9900k desktop system, CPU @3.60 GHz with 32 GB RAM and an NVIDIA GeForce RTX 2080 GPU. The performance of the model is evaluated on evaluation metrics i.e., ViT patch size, testing accuracy, precision, recall, and F1 score. The patch size of the ViT model is varied from smaller (14*14) pixels to larger (56*56) pixels to study its effect on the performance of the model. The number of epochs for training the model is considered 40 as there is no further increase in the accuracy and a decrease in the model loss after the 40th epoch. As shown in Table 1, the performance metrics of the ViT model are compatible and an accuracy of more than 80% is achieved for all patch sizes.

Table 1. Overall Results of the ViT model on various patch sizes

Patch Size	Testing Accuracy	Recall	Precision	F1 Score
14*14	0.8612	0.8840	0.8304	0.8564
21*21	0.8474	0.8613	0.8270	0.8438
28*28	0.8414	0.8518	0.8253	0.8383
56*56	0.8103	0.7896	0.8443	0.8161

Additionally, as shown in Figures 5, 6, 7, and 8, the training and validation curves (accuracy and loss) show slight divergence, which implies that the model has not been overfitted. Keeping the transformer patch size of 14*14, the ViT model achieved the highest testing accuracy, precision, recall, and F1 score of 0.8612, 0.8840, 0.8404, and 0.8564 respectively. Increasing the patch size to 21*21 pixels, the accuracy, precision, recall, and F1 scores decreased by 1.38, 2.27, 0.34, and 1.26% respectively. Further increasing the patch size to 28*28 pixels an accuracy of 0.8414, precision of 0.8518, recall of 0.8253, and F1 score of 0.8383 respectively. The lowest testing accuracy, precision, recall, and F1 score of 0.8103, 0.7896, 0.8443, and 0.8161 is recorded by the ViT model using the patch size of 52*52. The proposed model trained on a small patch size of 14×14 outperforms all the others in terms of all evaluation metrics. The confusion metrics of the model based on various patch size and number of parameters is depicted in Table 2.

The ViT model is then integrated with the sliding window approach for crack localization purposes in stone floor tiles. Testing images having a resolution of 1120*2240 acquired in various lightning conditions are taken. Using the sliding window approach, the test image is divided into 50 equal patches of size 224*224. Each patch is given to the trained ViT model to decide whether it belongs to the crack class or not. The patch classified as a crack is represented by a red bounding box as shown in Figure 9. The whole image and single patch inference time of the ViT model are recorded to be 4.624 sec and 0.092 sec respectively. The black boxes in Figure 9 represent the False Positives. The False negative patches did not exist as the system has the capability to correctly identified the crack region however the greater number of FP is due to the similarity of the grout lines to crack regions. The number of FP can be decreased by increasing the number of training data.

Table 2.Patch size Vs No of parameters and Performance

PS	NOP (Millions)	Confusion Matrices		
14*14	36.51	**Class**	**Crk**	**N-Crk**
		Crk	480	98
		N-Crk	63	519
21*21	16.10	**Class**	**Crk**	**N-Crk**
		Crk	478	100
		N-Crk	77	505
28*28	11.44	**Class**	**Crk**	**N-Crk**
		Crk	477	101
		N-Crk	83	499
56*56	**5.71**	**Class**	**Crk**	**N-Crk**
		Crk	480	90
		N-Crk	130	452

*PS: Patch Size, *NOP: Number of parameters, *Crk: Crack, *N-Crk: Non-Crack

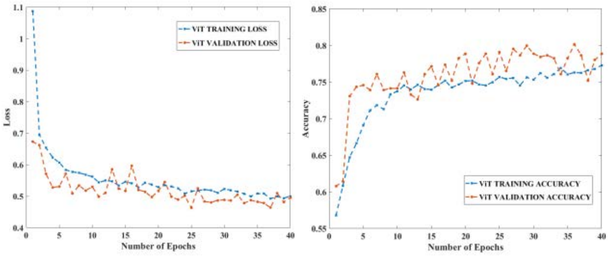

Figure 5. Training and validation (a) loss (b) Accuracy graphs of ViT model (14*14)

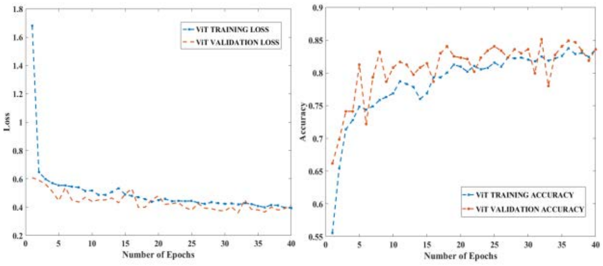

Figure 6. Training and validation (a) loss (b) Accuracy graphs of ViT model (21*21

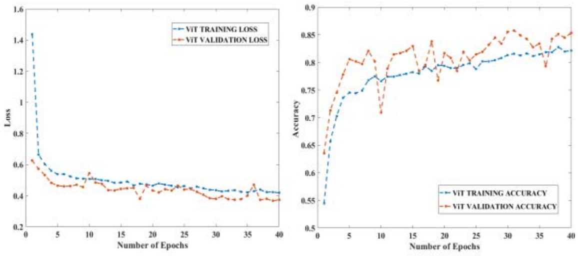

Figure 7. Training and validation (a) loss (b) Accuracy graphs of ViT model (28*28)

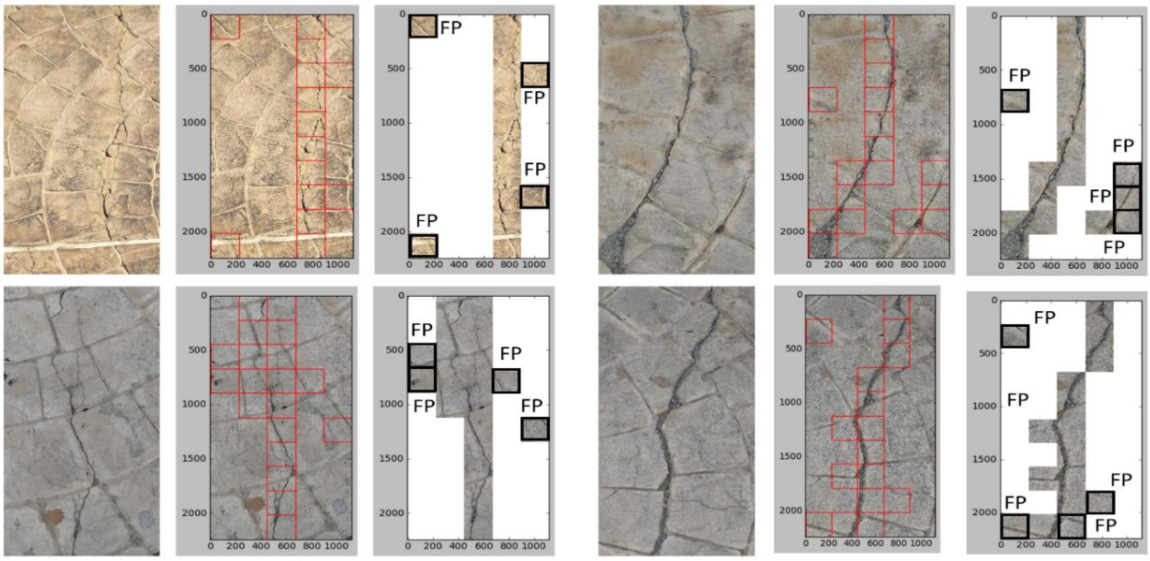

Figure 9. Crack localization and scanning for FP and FN using the sliding window approach.

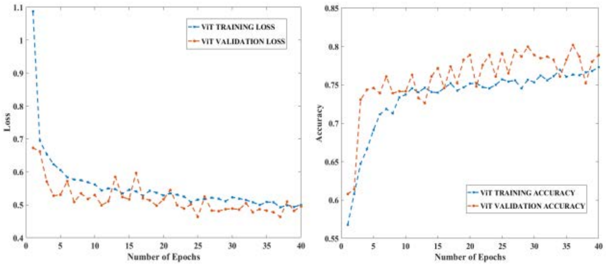

Figure 8. Training and validation (a) loss (b) Accuracy graphs of ViT model (56*56)

4 Discussion and Conclusion

This paper proposed a ViT-based framework for crack detection and localization in stone floor tiles. The performance of the proposed ViT model is compared to various input patch sizes. The experimental results showed that input patch size has a significant effect on the crack detection performance of the models. The model showed high crack detection performance on the lowest patch size. The performance of the model degraded as the patch size increased. It is also noted that decreasing the input patch increases the number of parameters of the ViT model which leads to an increase in the computational time and complexity of the model. The result of the crack localization in Figure 09 shows no FN and a small number of FP which shows that the model has the capability to localize the crack efficiently.

From the above discussion, it can be concluded that the ViT transformer integrated with the sliding window approach can be used to perform crack detection and localization in stone floor tiles. The ViT's ability to acquire global-scale features from the input image makes the task of crack detection. It can also be concluded that ViT transformers can be used for crack detection in various civil infrastructures i.e., pavement, concrete, bridges, and so on. Overall, the proposed ViT-based stone floor tiles crack detection method will enable Pavement inspection departments to automatically inspect the civil structure frequently. In the future, we are planning to add more data to the dataset to improve the accuracy of the proposed method.

References

[1] B. A. Graybeal, B. M. Phares, D. D. Rolander, M. Moore, and G. Washer, “Visual Inspection of Highway Bridges,” *Journal of Nondestructive Evaluation*, vol. 21, no. 3, pp. 67–83, Sep. 2002, doi: 10.1023/A:1022508121821.

[2] B. M. Phares, G. A. Washer, D. D. Rolander, B. A. Graybeal, and M. Moore, “Routine Highway Bridge Inspection Condition Documentation Accuracy and Reliability,” *Journal of Bridge Engineering*, vol. 9, no. 4, pp. 403–413, Jul. 2004, doi: 10.1061/(ASCE)1084-0702(2004)9:4(403).

[3] “Minneapolis Interstate 35W Bridge Collapse - Minnesota Issues Resources Guides.” https://www.lrl.mn.gov/guides/guides?issue=bridges (accessed Apr. 04, 2022).

[4] T. Asakura and Y. Kojima, “Tunnel maintenance in Japan,” *Tunnelling and Underground Space Technology*, vol. 18, no. 2, pp. 161–169, Apr. 2003, doi: 10.1016/S0886-7798(03)00024-5.

[5] I. Abdel-Qader, O. Abudayyeh, and M. E. Kelly, “Analysis of Edge-Detection Techniques for Crack Identification in Bridges,” *Journal of Computing in Civil Engineering*, vol. 17, no. 4, pp. 255–263, Oct. 2003, doi: 10.1061/(ASCE)0887-3801(2003)17:4(255).

[6] M. Kamaliardakani, L. Sun, and M. K. Ardakani, “Sealed-Crack Detection Algorithm Using Heuristic Thresholding Approach,” *Journal of Computing in Civil Engineering*, vol. 30, no. 1, p. 04014110, Jan. 2016, doi: 10.1061/(ASCE)CP.1943-5487.0000447.

[7] S. K. Sinha and P. W. Fieguth, “Morphological segmentation and classification of underground pipe images,” *Machine Vision and Applications*, vol. 17, no. 1, pp. 21–31, Apr. 2006, doi: 10.1007/s00138-005-0012-0.

[8] S. K. Sinha and P. W. Fieguth, “Automated detection of cracks in buried concrete pipe images,” *Automation in Construction*, vol. 15, no. 1, pp. 58–72, Jan. 2006, doi: 10.1016/j.autcon.2005.02.006.

[9] S. Chambon, P. Subirats, and J. Dumoulin, “Introduction of a wavelet transform based on 2D matched filter in a Markov Random Field for fine structure extraction: Application on road crack detection,” *Proceedings of SPIE - The International Society for Optical Engineering*, vol. 7251, Feb. 2009, doi: 10.1117/12.805437.

[10] Y. Fujita and Y. Hamamoto, “A robust automatic crack detection method from noisy concrete surfaces,” *Machine Vision and Applications*, vol. 22, no. 2, pp. 245–254, Mar. 2011, doi: 10.1007/s00138-009-0244-5.

[11] I. Abdel-Qader, S. Pashaie-Rad, O. Abudayyeh, and S. Yehia, “PCA-Based algorithm for unsupervised bridge crack detection,” *Advances in Engineering Software*, vol. 37, no. 12, pp. 771–778, Dec. 2006, doi: 10.1016/j.advengsoft.2006.06.002.

[12] X. Q. Zhu and S. S. Law, “Wavelet-based crack identification of bridge beam from operational deflection time history,” *International Journal of Solids and Structures*, vol. 43, no. 7, pp. 2299–2317, Apr. 2006, doi: 10.1016/j.ijsolstr.2005.07.024.

[13] X. Zhou, L. Xu, and J. Wang, “Road crack edge detection based on wavelet transform,” *IOP Conf. Ser.: Earth Environ. Sci.*, vol. 237, no. 3, p. 032132, Feb. 2019, doi: 10.1088/1755-1315/237/3/032132.

[14] L. Ali, F. Alnajjar, N. Zaki, and H. Aljassmi, “Pavement Crack Detection by Convolutional AdaBoost Architecture: 8th Zero Energy Mass Custom Home International Conference, ZEMCH 2021,” *ZEMCH 2021 - 8th Zero Energy Mass Custom Home International Conference, Proceedings*, pp. 383–394, 2021.

[15] L. Ali, S. Harous, N. Zaki, W. Khan, F. Alnajjar, and H. A. Jassmi, “Performance Evaluation of different Algorithms for Crack Detection in Concrete Structures,” in *2021 2nd International Conference on Computation, Automation and Knowledge Management (ICCAKM)*, Jan. 2021, pp. 53–58. doi: 10.1109/ICCAKM50778.2021.9357717.

[16] K. Chaiyasarn, W. Khan, L. Ali, M. Sharma, D. Brackenbury, and M. DeJong, “Crack Detection in Masonry Structures using Convolutional Neural Networks and Support Vector Machines,” Jul. 2018. doi: 10.22260/ISARC2018/0016.

[17] L. Ali, N. K. Valappil, D. N. A. Kareem, M. J. John, and H. A. Jassmi, “Pavement Crack Detection and Localization using Convolutional Neural Networks (CNNs),” in *2019 International Conference on Digitization (ICD)*, Nov. 2019, pp. 217–221. doi: 10.1109/ICD47981.2019.9105786.

[18] L. Ali, F. Sallabi, W. Khan, F. Alnajjar, and H. Aljassmi, “A deep learning-based multi-model ensemble method for crack detection in concrete structures,” *ISARC Proceedings*, pp. 410–418, Nov. 2021.

[19] L. Ali, F. Alnajjar, W. Khan, M. A. Serhani, and H. Al Jassmi, “Bibliometric Analysis and Review of Deep Learning-Based Crack Detection Literature Published between 2010 and 2022,” *Buildings*, vol. 12, no. 4, Art. no. 4, Apr. 2022, doi: 10.3390/buildings12040432.

[20] L. Ali, F. Alnajjar, H. A. Jassmi, M. Gocho, W. Khan, and M. A. Serhani, “Performance Evaluation of Deep CNN-Based Crack Detection and Localization Techniques for Concrete Structures,” *Sensors*, vol. 21, no. 5, Art. no. 5, Jan. 2021, doi: 10.3390/s21051688.

[21] F.-C. Chen and M. R. Jahanshahi, “NB-CNN: Deep Learning-Based Crack Detection Using Convolutional Neural Network and Naïve Bayes Data Fusion,” *IEEE Transactions on Industrial Electronics*, vol. 65, no. 5, pp. 4392–4400, May 2018, doi: 10.1109/TIE.2017.2764844.

[22] E. Asadi Shamsabadi, C. Xu, A. S. Rao, T. Nguyen, T. Ngo, and D. Dias-da-Costa, “Vision transformer-based autonomous crack detection on asphalt and concrete surfaces,” *Automation in Construction*, vol. 140, p. 104316, Aug. 2022, doi: 10.1016/j.autcon.2022.104316.

[23] L. Ali, H. A. Jassmi, W. Khan, and F. Alnajjar, “Crack45K: Integration of Vision Transformer with Tubularity Flow Field (TuFF) and Sliding-Window Approach for Crack-Segmentation in Pavement Structures,” *Buildings*, vol. 13, no. 1, Art. no. 1, Jan. 2023, doi: 10.3390/buildings13010055.

[24] S. Wang, X. Chen, and Q. Dong, “Detection of Asphalt Pavement Cracks Based on Vision Transformer Improved YOLO V5,” *Journal of Transportation Engineering, Part B: Pavements*, vol. 149, no. 2, p. 04023004, Jun. 2023, doi: 10.1061/JPEODX.PVENG-1180.

[25] K. Han *et al.*, “A Survey on Vision Transformer,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 87–110, Jan. 2023, doi: 10.1109/TPAMI.2022.3152247.

[26] A. Dosovitskiy *et al.*, “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.” arXiv, Jun. 03, 2021. doi: 10.48550/arXiv.2010.11929.

[27] J. Lee, J. Bang, and S.-I. Yang, “Object detection with sliding window in images including multiple similar objects,” in *2017 International Conference on Information and Communication Technology Convergence (ICTC)*, Oct. 2017, pp. 803–806. doi: 10.1109/ICTC.2017.8190786.

`