buildings

MDPI

Article

Detection of Bridge Damages by Image Processing Using the Deep Learning Transformer Model

Tomotaka Fukuoka and Makoto Fujiu *

Institute of Transdisciplinary Sciences for Innovation, Kanazawa University, Kanazawa 920-1192, Japan
* Correspondence: fujiu@se.kanazawa-u.ac.jp

Abstract: In Japan, bridges are inspected via close visual examinations every five years. However, these inspections are labor intensive, and a shortage of engineers and budget constraints will restrict such inspections in the future. In recent years, efforts have been made to reduce the labor required for inspections by automating various aspects of the inspection process. In particular, image processing technology, such as transformer models, has been used to automatically detect damage in images of bridges. However, there has been insufficient discussion on the practicality of applying such models to damage detection. Therefore, this study demonstrates how they may be used to detect bridge damage. In particular, delamination and rebar exposure are targeted using three different models trained with datasets containing different size images. The detection results are compared and evaluated, which shows that the detection performance of the transformer model can be improved by increasing the size of the input image. Moreover, depending on the target, it may be desirable to avoid changing the detection target. The result of the largest size of the input image shows that around 3.9% precision value or around 19.9% recall value is higher than one or the other models.

Keywords: bridge maintenance; damage detection; image size; transformer model

check for updates

Citation: Fukuoka, T.; Fujiu, M. Detection of Bridge Damages by Image Processing Using the Deep Learning Transformer Model. *Buildings* **2023**, *13*, 788. https://doi.org/10.3390/buildings13030788

Academic Editors: Zhigang Shen, Endong Wang, Ri Na, Zhexiong Shang and Chongsheng Cheng

Received: 31 January 2023
Revised: 13 March 2023
Accepted: 14 March 2023
Published: 16 March 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In Japan, there are approximately 730,000 road bridges that are at least 2 m in length. Most of these bridges were constructed during the high-economic-growth period; consequently, the proportion of aged bridges is increasing rapidly and their maintenance, management, and renewal must be planned carefully [1]. Effective bridge management includes preventative maintenance, where measures are taken to address minor damage before it progresses; however, this requires inspections to be conducted periodically. Since 2014, road administrators have been obliged to conduct 100% surveillance of bridges via close visual inspection once every five years, according to the unified standard established by the government [2]. However, such inspections are costly, and the financial burden will present challenges in the future; therefore, various methods have been considered to improve efficiency and reduce the cost of inspections.

For example, image-processing technology has been considered to detect damage automatically. Currently, 26 types of damage are considered during bridge inspections, which are mainly confirmed visually [2]; therefore, it should be possible to detect such damage using image-processing techniques. Moreover, deep-learning techniques have been proposed to automatically detect cracks [3–5], corrosion, delamination/rebar exposure, and water leakage [6,7]. In some cases, such as large structures, a drone is used when it is difficult to obtain an overall picture of the structure from the ground alone [8]. Two methods are typically used to detect damage via deep learning: object detection and semantic segmentation where the outputs consist of a rectangle containing the target and the pixels containing the target, respectively. Object detection methods could detect objects quickly from an input image [9,10]. However not suitable for accurate nonlinear damage area acquisition. A semantic segmentation method can accurately detect nonlinear objects

such as humans and animals, and vehicles [11,12]. Furthermore, transformer models [13], which are commonly used in the field of natural language processing, have been applied to image processing in recent years [14]. In natural language processing, a word string from a sentence is used as the input for the transformer model, and the position of the word within the sentence provides context. When applied to image processing, the sequence of images obtained by dividing an original image is treated as a word, and the position of the divided image within the original image provides context for processing. CNN-based detection models extract and process the entire feature set of the input image; Transformer-based models also extract features of the input image but can consider relationships between each segmented input image.

In general, deep learning requires a large amount of training data to demonstrate the performance of a model. The amount of training data can be increased by augmenting datasets with rotated or flipped images, and this technique is widely used in current research. However, the quality of the image data and the class ratio should also be considered.

When a deep-learning model is based on images that have a size different from the actual input images, then preprocessing steps such as image enlargement, reduction, or cropping are required. However, these techniques change the amount of information that the model references during the detection process; for example, features in the image may change or the area of the input image may be smaller than that of the original image. Therefore, the effects of preprocessing on detection accuracy must be considered. Thus far, there has been insufficient discussion of these issues concerning bridge damage detection using transformer models.

This study focused on the detection of bridge damage in the form of delamination and rebar exposure via image processing using a transformer model. Note that delamination in Japanese bridge inspections refers to the delaminated surface of concrete members. The effect of the sizes of the images used to train the model and the size of the input image at the time of detection were investigated. The detection result consisted of a pixel-by-pixel damage detection image obtained via semantic segmentation using detection models trained with different size images. This is because the location, size, and shape of the damage must be considered in the actual bridge inspection to assess the health of the bridge. The novelty of this study is the evaluation of the effectiveness of the size of the image for the detection of bridge damage.

2. Related Research

In recent years, many new inspection methods have been proposed to reduce economic costs and simplify the process [15–17]. Especially, deep learning-based methods have been studied well. For example, research is being conducted on the automatic diagnosis of buildings and the automatic generation of inspection results [18,19].

In semantic segmentation, an object within an image is identified on a pixel-by-pixel basis according to a pre-trained detection object [11,12,20]. The results consist of a pixel unit and they can be combined with data from the time of the shooting, such as the shooting distance and focal length, to determine information such as the location and dimensions of any damage. The information required to determine the degree of damage can also be obtained.

Various semantic segmentation methods using transformer models have been proposed [21–23]. These studies have shown that methods using transformer models can conduct classification with greater accuracy than those using convolutional learning techniques, which have been widely used. Models, where the damage-detection target is limited, have also been proposed. For example, Liu et al. proposed a semantic segmentation model, CrackFormer [24], for detecting cracks and compared it with existing models. They showed that the CrackFormer model performed better than existing models. However, existing studies have not considered the size of the image data during training, nor have they evaluated the effects of the image size on detection.

3. Materials and Methods

In this study, we trained a detection model to identify delamination and rebar exposure using transformer models for datasets with different image sizes. We compared the detection performance of these models and evaluated the effects of the image size. Furthermore, for comparison with existing methods, we compared the detection results with those obtained using convolutional-learning models.

3.1. Deep-Learning Model

In this study, we conducted semantic segmentation using the transformer model, SegFormer [25], which has shown good results in the segmentation of various objects. For comparison, we also used the convolutional-learning model, SegNet [12]. Previous studies have shown that a larger encoder part improves the classification performance of SegFormer; therefore, we used the SegFormer B-5 model, which has the highest classification performance. According to the paper, we train the SegFormer model with settings that the optimization algorithm is adamW, the learning rate is 0.00006, and the number of the epoch is 20. Figure 1 shows the accuracy and the loss of each epoch during model training with dataset A will be described in Section 3.2. Similarly, the SegNet model also would be trained with settings that the optimization algorithm is adam, the learning rate is 0.0001, the parameter is 0.9, is 0.999, 4000 iterations per epoch, the size of the mini-batch is 16, and the number of the epoch is 100. Each model is trained with the dataset of the pair of RGB images and annotated images of the target damage place. The SegFormer model assumed that the input images were 512 × 512 pixels and the SegNet model assumed that the input images were 224 × 224 pixels. The SegFormer model was trained using the large-scale image dataset ImageNet, which was classified into 1000 classes, and then used to train a detection model for delamination and rebar exposure. Similarly, for the SegNet model, the VGG16 model was trained using ImageNet and then used to train a detection model for delamination and rebar exposure.

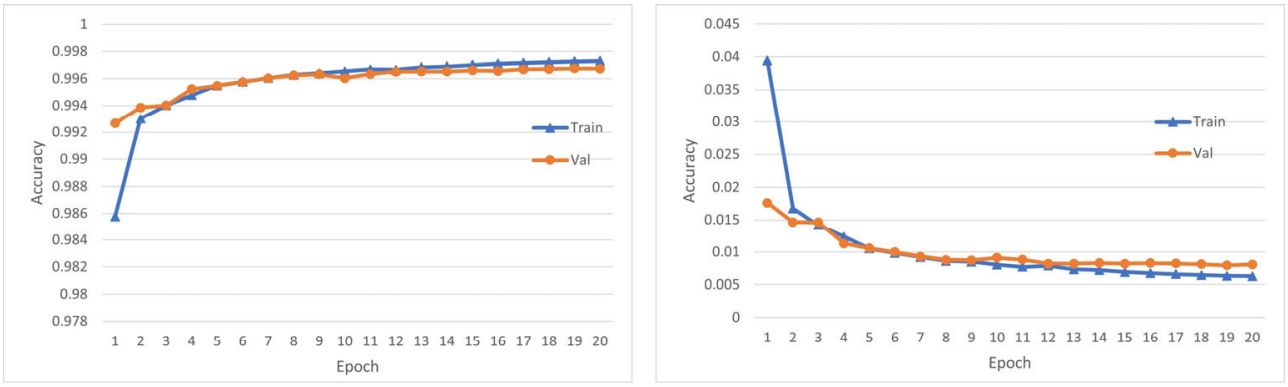


Figure 1. Example of the accuracy and the loss of each epoch during model training. (**a**) accuracy, (**b**) loss.

3.2. Dataset

In this study, 179 images from the bridge inspection report for K prefecture in Japan were used as images of delamination and rebar exposure. These images were taken of portions of the actual bridge to visually document the damage identified during the inspection and include specific damage. This dataset was annotated by a single annotator, with areas of delamination and rebar exposure marked in black and red, respectively. The image size and aspect ratio were not standardized, the maximum length of the long side was 950 pixels, and the minimum length of the short side was 270 pixels. The dataset was divided into training and testing datasets, which contained 119 and 60 images, respectively. The training dataset was augmented to increase the number of training images. The augmentation

consisted of the following processes conducted in random order: horizontal flipping, where the image was flipped on a horizontal axis; vertical flipping, where the image was flipped on a vertical axis; translation, where the image was randomly translated by 0–20 pixels horizontally and 0–20 pixels vertically; scaling, where the image was randomly scaled by a factor of 0.8–1.2 horizontally and 0.8–1.2 vertically; and rotation, where the image was rotated randomly by 0–360°. It was assumed that the probability of each processing step not being conducted was 66% (Figure 2).

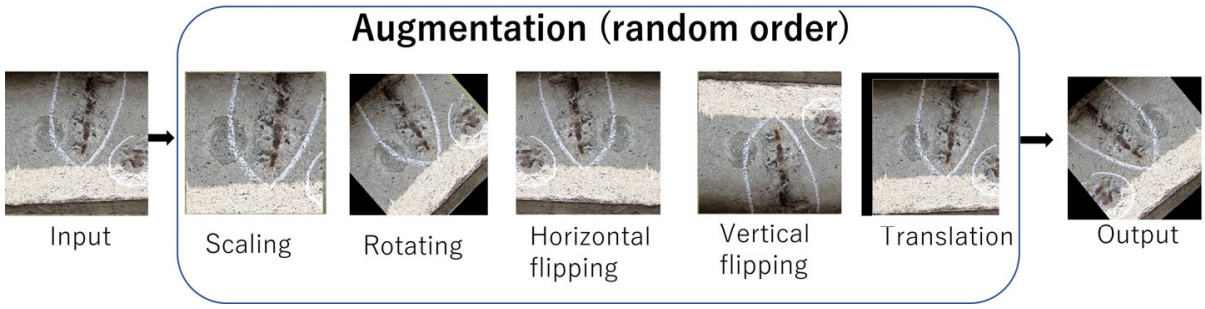


Figure 2. Example of augmentation.

To evaluate the effects of image size on performance, we created three training datasets: dataset A, with no additional processing; dataset B, where each image in the original dataset was divided into a mesh 224 pixels wide; and dataset C, where each image in the original image dataset was divided into a mesh 448 pixels wide (Figure 3). The data in dataset A were adjusted for size and aspect ratio when they were input into the model, whereas those in datasets B and C were only adjusted for size. For example, the image in dataset B is stretched from 224 pixels to the model's input size of 512 pixels upon entry into the model. Figure 4 shows an example of training data, a pair of images, and label data from dataset A. Figure 5 shows an example of training data, and a pair of images and label data from dataset B. Figure 6 shows an example of test data. Table 1 lists the number of images in the divided datasets.

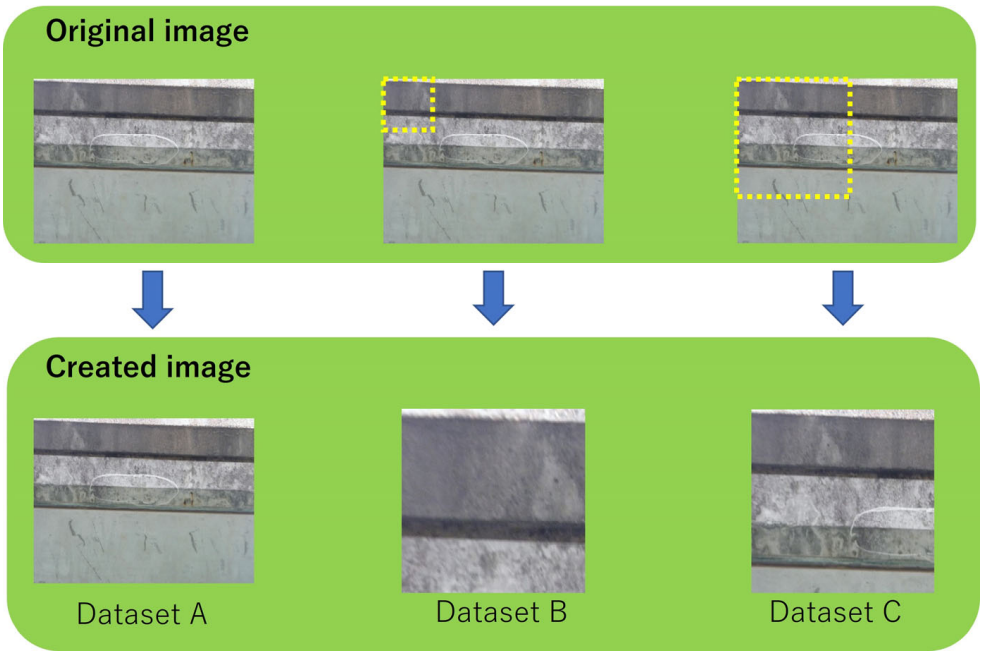


Figure 3. Example of each dataset image created by the original image.



Figure 4. Example of a pair of images and label data from dataset A.



Figure 5. Example of a pair of images and label data from dataset B.



Figure 6. Example of test data.

Table 1. Number of images in the datasets after division.

	Dataset B	Dataset C
Number of images after division	1718	478
Number of annotated images	369	464

The areas of delamination and the rebar exposure of the image of the training dataset are part of the image; therefore, the image may not be included when the image is divided; however, those images were used as training data, as shown in Table 1. Each dataset was padded after division, and the number of images in each dataset was increased to 17,810 images.

3.3. Detection Method

For each model, the input images were preprocessed by transforming them to match the size of the training images. No preprocessing was conducted during detection by model A, which was trained using dataset A. During detection by models B and C, which were

trained using datasets B and C, respectively, the input images were divided into meshes of 224 and 448 pixels, respectively. The divided images were received as input data, the detection results were output for each image, and they were recombined to generate the detection result for the original input image. This process is illustrated in Figure 7.

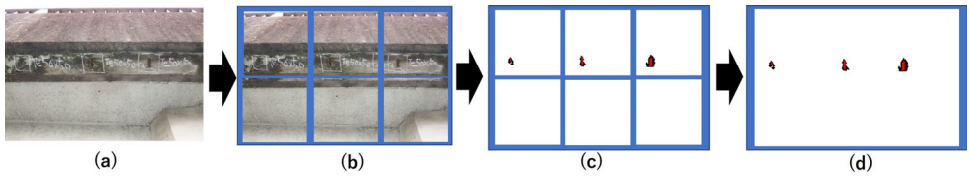


Figure 7. Example of the input and output of the image segmentation model. (**a**) Input image, (**b**) input image after division, (**c**) detection result of the divided image, and (**d**) final output result.

Dividing the images for models B and C implied that the range of the input image was narrower than the original image. To compensate for this, we overlayed the results from four sets of divided images, where the starting points of the divided images were shifted between the sets. The results of overlaying the following four detection results were set as the final detection result, that is, the starting points of the divisions were (1) not shifted, (2) shifted in the x-direction by half the width of the divided image, (3) shifted in the y-direction by half the height of the divided image, and (4) shifted in the x- and y-directions by half the width and height of the divided image (Figure 8). The result obtained by overlaying the four detection results was the final detection result (Figure 9).

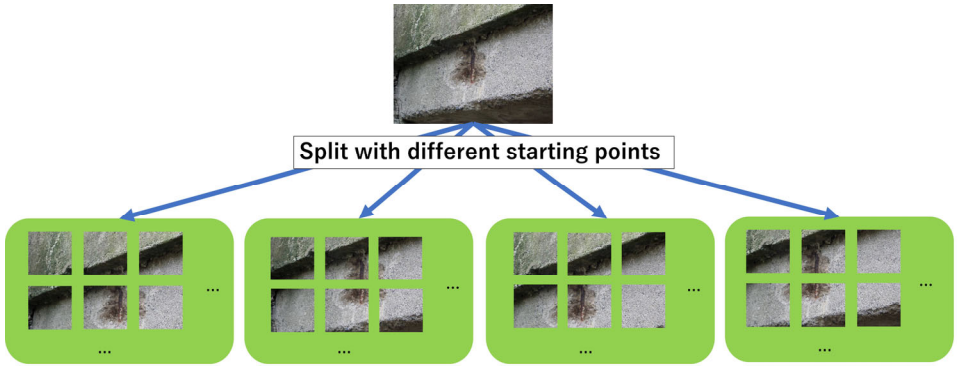


Figure 8. Example of the split images of four different starting points.

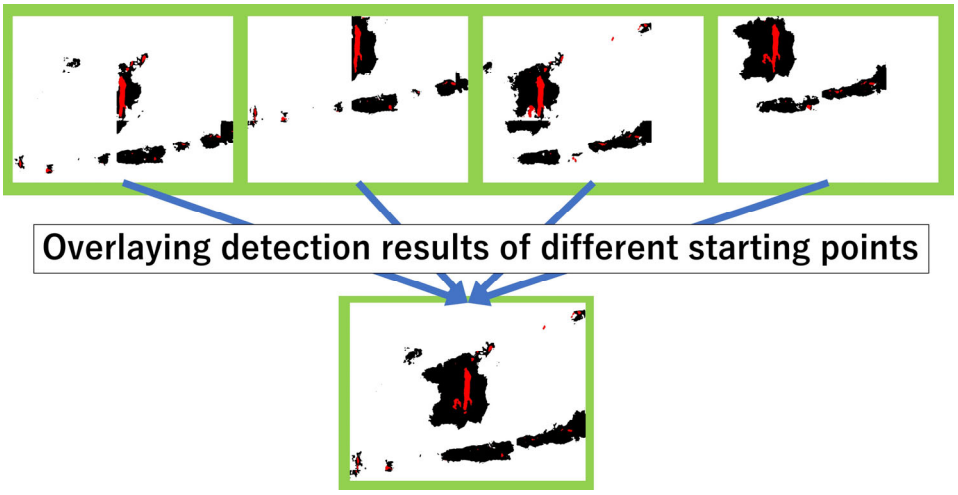


Figure 9. Overlay image of detection results for each division starting point.

3.4. Evaluation Indices

Precision, recall, and F-measure were used to evaluate the models in this study. They were calculated based on the success or failure of detection in pixel units based on the following example 1 to 3 of an equation:

$$\text{Precision} = \frac{\text{N}_{\text{much}}}{\text{N}_{\text{detect}}}, \tag{1}$$

and

$$\text{Recall} = \frac{\text{N}_{\text{much}}}{\text{N}_{\text{damage}}}, \tag{2}$$

and

$$\text{F} - \text{measure} = 2\frac{\text{Precision Recall}}{\text{Precision} + \text{Recall}}. \tag{3}$$

Here, N_{much} denotes the number of true positive pixels that the model correctly identified as the target (damage), N_{detect} is the total number of true positive/false positive pixels that the model identified as the target, and N_{damage} is the total number of true positive/false negative pixels belonging to the target.

Herein, precision refers to how accurately only the detection target damage pixels are detected as detection target damage among the pixels assessed to be detection target damage in the detection results.

In other words, a higher precision indicates higher accuracy in the model detecting only the detection target damage. Recall indicates how many of the correct detection target damage pixels were detected by the model. A higher recall indicates fewer missed cases of target damage detection. F-measure is a harmonic mean of precision and recall.

4. Results

The delamination and rebar-exposure detection results obtained using SegFormer trained on datasets A, B, and C (labeled SF_A, SF_B, and SF_C, respectively), and the comparison model SegNet trained on dataset B (labeled SN_B) are listed in Tables 2 and 3.

Table 2. Comparison of the delamination detection results.

	SF_A	SF_B	SF_C	SN_B
Precision	0.808	0.771	0.769	0.793
Recall	0.741	0.658	0.674	0.441
F-measure	0.773	0.710	0.718	0.567

Table 3. Comparison of the rebar-exposure detection results.

	SF_A	SF_B	SF_C	SN_B
Precision	0.708	0.747	0.765	0.693
Recall	0.685	0.486	0.563	0.420
F-measure	0.696	0.589	0.649	0.523

As shown in Table 2, SF_A had the highest precision and recall and F-measure values. The other SegFormer models had slightly lower precision than the SegNet model; however, they had much better recall. Therefore, the transformer models have high F-measure values and are effective for detecting delamination. Comparing the results for SF_B and SF_C shows that larger divisions resulted in better detection performance. Furthermore, comparing the results for SF_A and SF_B shows that SF_A had better detection performance. When comparing all the SegFormer results, in the case of delamination detection, increasing the range of the

input image improved the detection performance, whereas adjusting the aspect ratio had little effect.

As shown in Table 3, the SegFormer model demonstrated better detection performance than the SegNet model, and the transformer model was effective for detecting exposed rebar. As with delamination, the recall and F-measure of SF_A, SF_B, and SF_C improved as the range of the input image increased; however, SF_C had the highest precision. This is probably because, in addition to having a wide range, it is beneficial if the aspect ratio is not adjusted so that information regarding the shape of the rebar is preserved.

Overall, these results suggest that, for the detection of delamination and rebar exposure, the performance of the transformer model can be improved by increasing the pixel size of the input image. Moreover, because the precision decreased when detecting rebar exposure, it is desirable to avoid changing the aspect ratio of the dataset depending on the target to suppress excessive detection. Finally, when the input image was smaller than the assumed input size of the detection model, the precision fluctuated by −5.7–3.9% and the recall decreased by 19.9%, and the F-measure decreased by 10.7%; therefore, this should be avoided. On the other hand, the results of this evaluation are not accurate enough to replace humans. A more accurate method is needed to reduce the cost of current handcraft inspection.

5. Conclusions

In recent years, it has become necessary to develop an alternative to the close visual inspection of bridges, and a variety of damage-detection methods using image processing have been proposed. This study considered damage detection using a transformer model, which is the latest deep-learning-based detection model, and we evaluated an adaptive method focused on image size. We compared the results obtained when input images of different sizes were used during training and testing, and showed that the image size affected the detection performance. The detection model trained on the largest image size data set performed better than other models trained on smaller image sizes. However, the improvement effect depends on the target damage.

Camera performance has been improving year by year, and in recent years, images with high pixel counts of about 4K can be easily acquired. Therefore, the number of pixels in images acquired during the actual inspection is also expected to increase. In the future, we will evaluate images that are larger than the size of the model to investigate the upper limit where the image size works effectively for detection. In addition, since the number of original image datasets used in this paper is very small, we will also evaluate a larger dataset to assess the effect of the size of the dataset. The validity of image size should also be evaluated for other types of damage such as cracks, rust, etc.

Author Contributions: Methodology, T.F.; Writing—original draft, T.F.; Project administration, M.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data are not publicly available due to it is not open data.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ministry of Land, Infrastructure, Transport and Tourism. White Paper. 2021. Available online: https://www.mlit.go.jp/hakusyo/mlit/r02/hakusho/r03/pdf/kokudo.pdf (accessed on 30 September 2022).
2. Ministry of Land, Infrastructure, Transport and Tourism. Road Bridge Periodic Inspection Procedures, Road Bureau. 2022. Available online: https://www.mlit.go.jp/road/sisaku/yobohozen/tenken/yobo4_1.pdf (accessed on 30 September 2022).
3. Chun, P.-J.; Igo, A. Crack detection from image using Random Forest. *J. Jpn Soc. Civ. Eng. F3* **2015**, *71*, 1–8.
4. Yokoyama, S.; Matsumoto, T. Development of an automatic detector of cracks in concrete using machine learning. *Procedia Eng.* **2017**, *171*, 1250–1255. [CrossRef]
5. Cha, Y.-J.; Choi, W.; Büyüköztürk, O. Deep learning-based crack damage detection using convolutional neural networks. *Comput. Aided Civ. Infrastruct. Eng.* **2017**, *32*, 361–378. [CrossRef]

6. Dong, H.; Gang, T. Damage detection of quayside crane structure based on improved faster R-CNN. *Int. J. New Dev. Eng. Soc.* **2019**, *3*, 284–301.
7. Zhang, C.; Chang, C.; Jamshidi, M. Concrete bridge surface damage detection using a single-stage detector. *Comput. Aided Civ. Infrastruct. Eng.* **2020**, *35*, 389–409. [CrossRef]
8. Bianchi, E.; Abbott, A.L.; Tokekar, P.; Hebdon, M. COCO-bridge: Structural detail data set for bridge inspections. *J. Comput. Civ. Eng.* **2021**, *35*, 04021003. [CrossRef]
9. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
10. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. SSD: Single Shot Multibox Detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
11. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.
12. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: Deep convolutional encoder-decoder architecture for image segmentation:1511.00561v3. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef] [PubMed]
13. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, N.A.; Kaiser, Ł.; Polosukhin, I. Attention is All you Need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 6000–6010.
14. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. In Proceedings of the International Conference on Learning Representations, Virtual, 3–7 May 2021.
15. Menendez, E.; Victores, J.G.; Montero, R.; Martínez, S.; Balaguer, C. Tunnel structural inspection and assessment using an autonomous robotic system. *Autom. Constr.* **2018**, *87*, 117–126. [CrossRef]
16. Pham, N.H.; La, H.M.; Ha, Q.P.; Dang, S.N.; Vo, A.H.; Dinh, Q.H. Visual and 3D Mapping for Steel Bridge Inspection using a Climbing Robot. In Proceedings of the ISARC 2016—33rd International Symposium on Automation and Robotics in Construction, Auburn, AL, USA, 18–21 July 2016; pp. 141–149.
17. Xie, R.; Yao, J.; Liu, K.; Lu, X.; Liu, Y.; Xia, M.; Zeng, Q. Automatic multi-image stitching for concrete bridge inspection by combining point and line features. *Autom. Constr.* **2018**, *90*, 265–280. [CrossRef]
18. Rui, Z.; Ruqiang, Y.; Zhenghua, C.; Kezhi, M.; Peng, W.; Robert, X.G. Deep learning and its applications to machine health monitoring. *Mech. Syst. Signal Process.* **2019**, *115*, 213–237.
19. Soleimani-Babakamali, M.H.; Esteghamati, M.Z. Estimating seismic demand models of a building inventory from nonlinear static analysis using deep learning methods. *Eng. Struct.* **2022**, *266*, 114576. [CrossRef]
20. Li, S.; Zhao, X.; Zhou, G. Automatic pixel-level multiple damage detection of concrete structure using fully convolutional network. *Comput. Aided Civ. Infrastruct. Eng.* **2019**, *34*, 616–634. [CrossRef]
21. Jin, Y.; Han, D.; Ko, H. TrSeg: Transformer for semantic segmentation. *Pattern Recognit. Lett.* **2021**, *148*, 29–35. [CrossRef]
22. Xu, Z.; Zhang, W.; Zhang, T.; Yang, Z.; Li, J. Efficient transformer for remote sensing image segmentation. *Remote Sens.* **2021**, *13*, 3585. [CrossRef]
23. Li, Z.; Xu, P.; Xing, J.; Yang, C. SDFormer: A novel transformer neural network for structural damage identification by segmenting the strain field map. *Sensors* **2022**, *22*, 2358. [CrossRef] [PubMed]
24. Liu, H.; Miao, X.; Mertz, C.; Xu, C.; Kong, H. CrackFormer; Transformer Network for Fine-Grained Crack Detection. In Proceedings of the Institute of Electrical and Electronics Engineers/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 3763–3772.
25. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, M.J.; Luo, P. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. *arXiv* **2021**, arXiv:2105.15203.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.