Classification Model Based on U-Net for Crack Detection from Asphalt Pavement Images

Yusuke Fujita*, Taisei Tanaka, Tomoki Hori, and Yoshihiko Hamamoto

Graduate School of Sciences and Technology for Innovation, Yamaguchi University, Ube, Japan
*Correspondence: y-fujita@yamaguchi-u.ac.jp (Y.F.)

***Abstract*—The purpose of our study is to detect cracks accurately from asphalt pavement surface images, which includes unexpected objects, non-uniform illumination, and irregularities in surfaces. We propose a method to construct a classification Convolutional Neural Network (CNN) model based on the pre-trained U-Net, which is a well-known semantic segmentation model. Firstly, we train the U-Net with a limited amount of the asphalt pavement surface dataset which is obtained by a Mobile Mapping System (MMS). Then, we use the encoder of the trained U-Net as a feature extractor to construct a classification model, and train by fine-tuning. We describe comparative evaluations with VGG11, ResNet18, and GoogLeNet as well-known models constructed by transfer learning using ImageNet, which is a large size dataset of natural images. Experimental results show our model has high classification performance, compared to the other models constructed by transfer learning using ImageNet. Our method is effective to construct convolutional neural network model using the limited training dataset.**

***Keywords*—Convolutional Neural Network (CNN), crack detection, pavement inspection, Mobile Mapping System (MMS), pre-training, U-Net**

I. Introduction

Application of machine vision is expected for efficiency and objectivity of inspection in various fields. Automation of visual inspection for asphalt pavement surface images is also expected, but it is difficult to detect cracks with high accuracy, because there are unexpected objects, non-uniform illumination, and irregularities in the pavement surface.

Recently, a large number of methods are proposed to overcome these problems, which include many methods using convolutional neural network (CNN) models [1–4]. In general, a large amount of training dataset is needed to construct a high accurate CNN model. Especially, asphalt pavement surface images acquired during the daytime contain non-uniform and irregularly illuminated conditions. Additionally, annotated data is also needed for training dataset. Especially, in annotation for semantic segmentation, it is required to trace cracks at pixel level accurately. It needs very high cost for inspection specialists.

For practical application of this technique in the field, a method to establish a model with a limited number of small samples is expected. Major approaches of using CNN with small training data are to be applied transfer learning [5], in which firstly a CNN model trained using ImageNet dataset and fits to the target dataset by fine tuning [6]. Many studies have shown its effectiveness in various applications. However, because ImageNet is natural image dataset, the domain gap between natural images and asphalt pavement surface images as target domain may result in poor model performance.

In this article, we propose a method to construct a classification CNN model using a limited size of dataset. To summarize, the main contributions of our work are as follows:

- We propose a method for construction of a classification CNN model based on the trained U-Net model [7], which was the semantic segmentation model trained with a limited amount of training data. Our model's aim is to classify if cracks exist or not in each small region.
- We conducted comparative evaluation with some of the widely used CNN models, which were trained by transfer learning using ImageNet dataset.
- Experimental results show that our model has high classification performance, compared to well-known models constructed by transfer learning using ImageNet.

The rest of this article is structured as follows. Section II introduces related works, and Section III describes the proposed method in detail. Experiments and results are shown in Section IV and V. Finally, the conclusions are presented in Section VI.

II. Related Work

Many researchers have proposed various methods for defect detection and evaluation from surface images of infrastructures [1–4, 8, 9]. There are various methods based on image processing techniques and machine learning methods. Especially, many methods using deep learning techniques have been proposed.

Manuscript received August 5, 2022; revised September 5, 2022; accepted October 22, 2022.

Chun and Yamane *et al.* [8] proposed a method based on deep learning for automatic crack detection in asphalt pavement. In the article, divided images were classified to 6 classes, which include road markings, utility holes or bridge joints, to improve classification accuracy. Here, 30,000 of images or more were used to train the CNN model. In general, a large amount of training dataset is needed to construct a high accurate CNN model.

Under condition using a limited training dataset, transfer learning with a large amount of natural image dataset, such as ImageNet, is applied in various fields, and its effectivity has been reported. Liu *et al.* [9] also used deep learning and infrared thermography for asphalt pavement crack severity classification. The dataset of asphalt pavement crack was built in this work, including four levels of crack severity, no crack, low-severity crack, medium-severity crack, and high-severity crack. 13 typical CNN models, which include eight pre-trained CNN models trained by ImageNet for transfer learning, were trained and evaluated. However, to handle many classes, more training data is needed to train the CNN model. In addition, ImageNet dataset is natural images. The domain gap between natural images and asphalt pavement surface images may result in poor model performance. It is expected to improve performance of CNN models using the limited practical image dataset.

On the other hand, there are several types of CNN models that deal with different tasks, that are classification, object detection, and semantic segmentation. U-Net [7] is a well-known FCN (fully convolutional network) based model, which is proposed by Ronneberger *et al.* for efficiently segmenting biological microscopy images. The U-Net architecture comprises two parts, a contracting path to capture context, and a symmetric expanding path that enables precise location [10].

III. Method

A. Overview

Fig. 1 shows overview of the proposed method. Firstly, we build the U-Net model as a segmentation model, which model extract crack pixels on pixel level coarsely. After that, we use the encoder of the trained U-Net model as feature extractor to build our classification model. Our model trained by fine-tuning using training data for classification, in where the weights of feature extraction layers pre-trained on the U-Net are also updated. Finally, our model trained by two steps described above, classifies if cracks exist in the small region which is divided from the asphalt pavement surface images.

B. U-Net

Firstly, the U-Net model as a semantic segmentation model is trained using small dataset. Here, because it is difficult to construct a segmentation model with high accuracy, the strategy of this procedure is to construct a feature extractor which can extract richer information from cracks in asphalt pavement surface images, than that of the classification model trained simply.

Generally, for training semantic segmentation model such as U-Net, annotation which is labeling each pixel is very costly. Additionally, it is difficult to annotate the label of each pixel, especially on the boundary pixels of crack and non-crack regions, because of the inherent ambiguity of these pixels in semantic segmentation. Label ambiguity can also happen if we are not confident in the labels we provide for an image.

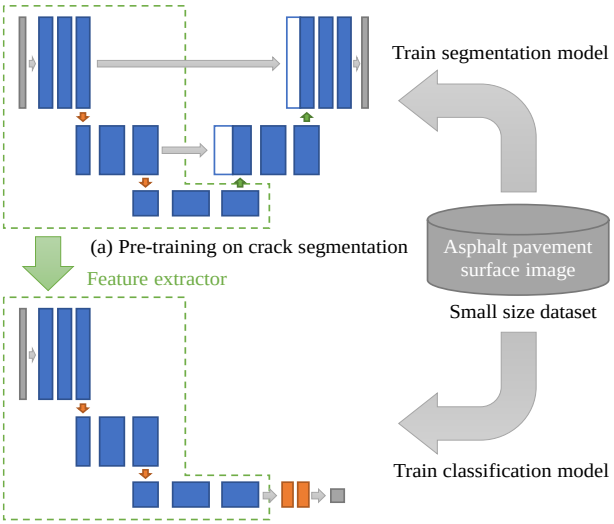


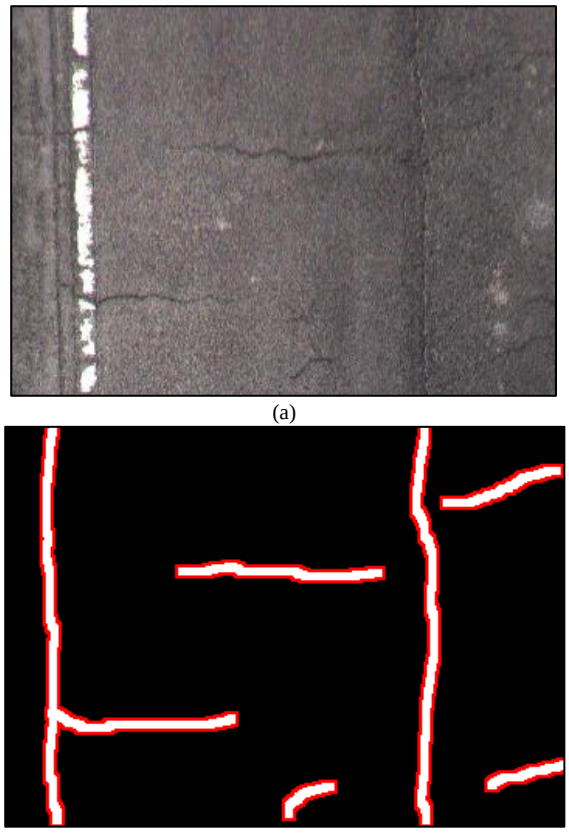
Figure 1. Overview of the proposed method to construct a classification CNN model using a limited amount of training data. The encoder of the trained U-Net is used as the feature extractor in our proposed classification model.

Our strategy to overcome the problem of label ambiguity is to solve three class segmentation, which classes are crack, background, and boundary class. The boundary class pixels are located between the crack and background classes, that can be difficult to assign to either of the two classes. Fig. 2 shows an example of annotated images. From the example shown in Fig. 2, the boundary between crack and background regions are difficult to label accurately. These ambiguities of these pixels may lead stagnation in learning process. In additionally, the number of pixels belonging the crack class is very small compared to that of the background class. In training the model using the unbalanced data, it is likely to be lost classification accuracy of small regions such as cracks by the trained model to reduce the estimated loss function. To solve this problem effectively, we also adjust the weight for each class used in calculation of training and validation loss, to avoid stagnation of training segmentation model.

Classification Model Based on U-Net

Our classification CNN model proposed here consists of the feature extractor trained in the U-Net model described above and the fully connected layers which is used as classification unit. The feature extractor trained using not a large amount of dataset such as ImageNet but asphalt pavement surface image data directly, to extract valuable feature for crack detection effectively. While the trained U-Net may not have the enough high segmentation performance, it is expected to construct a good feature

extractor to detect cracks during training U-Net. Then our model trained by fine-tuning using training data for classification, in where the weights of feature extraction layers pre-trained on the U-Net are also updated.



(a)

(b)

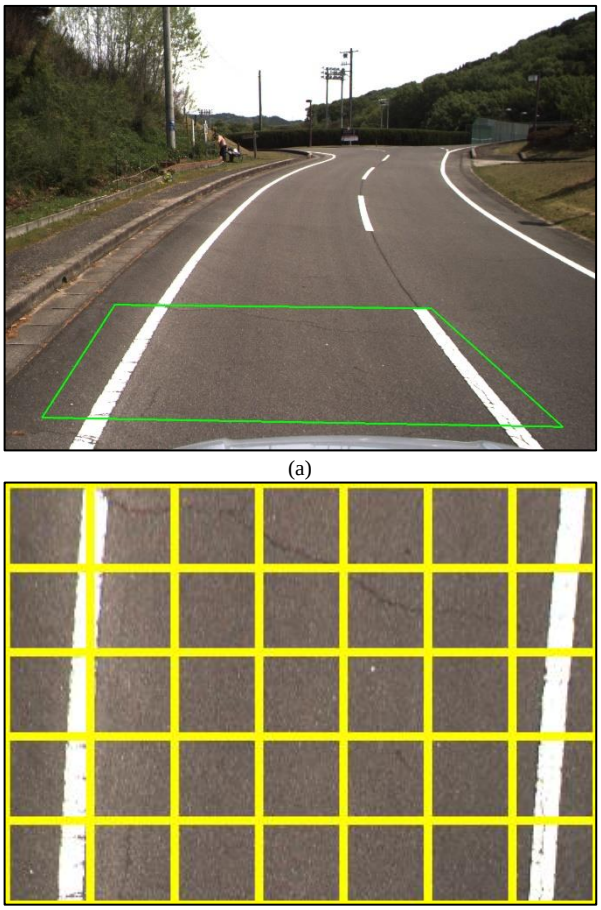
Figure 2. Perspective transformed image obtained by a Mobile Mapping System (MMS) and annotation data. Cracks in the perspective transformed image is traced manually with 3-pixel width. The boundary pixels between crack and background are additional class for semantic segmentation of our problem.

C. Experiments

We conducted experiments to exam if our pre-training U-Net is effective to build the classification CNN model which contains the encoder of pre-trained U-Net as the feature extractor. Additionally, in our comparative evaluation, VGG11 [11], ResNet18 [12], and GoogLeNet [13] as well-known models using a large amount of dataset ImageNet are also evaluated to compare our model. These models are compared with and without pre-trained using ImageNet dataset, to confirm effect of pre-training using ImageNet dataset.

A. Dataset and Settings

260 actual images of asphalt pavement surface were used in our experiment. 200 images and 60 images were acquired on different routes on different days. These are used for training and test separately. Fig. 3 shows the example of the asphalt pavement surface images and the perspective transformed image. The images were obtained using a Mobile Mapping System (MMS). MMS consists of a vehicle-mounted GPS antenna, laser scanners, cameras, and other equipment, which enables the efficient acquisition of highly accurate 3D positional information such as road contours while driving. Sketch images created manually by hand were used as the ground truth in quantitative evaluation. The cracks in the images are traced manually with 3-pixel width lines at pixel level, to be used for training and evaluation of both segmentation and classification performance. Firstly, the images were perspective transformed to orthographic projection images. The perspective transformed images have 250×350 resolution. The size of each pixel is approximately 1cm×1cm. In our evaluation of the classification performance, regions which contain over 30 crack pixels are used as crack class and the other regions are used as non-crack class.



(a)

(b)

Figure 3. Example of asphalt pavement surface data. Image (a) shows an original image obtained by a Mobile Mapping System (MMS). The green line indicates the target region of inspection. The horizontal length of the image is 350 cm, and the vertical length is 250 cm. Image (b) shows the perspective transformed target region. In actual visual inspection, the whole area is divided to 5×7 regions marked with yellow rectangles, which are classified into clacked region or non-cracked region.

B. Evaluation Metrics

We evaluated classification performance of the proposed CNN model using recall, precision, and F1-score. Recall and precision are calculated as:

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}, \quad (1)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}, \quad (2)$$

where TP, FN and FP denote true positive, false negative, and false positive, respectively. TP means the rate of correctly detected cracked regions in the truth cracked regions. FN means the rate of the missed crack regions in the truth non-cracked regions. FP also means the rate of the over-detected crack regions in the truth non-cracked regions. F1-score is a combined version of recall and precision. F1-score is defined as the harmonic mean of recall and precision.

$$\text{F1-score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision+Recall}} = \frac{2 \cdot \text{TP}}{2 \cdot \text{TP+FP+FN}}. \quad (3)$$

C. Implementation Details

We implemented our method and comparative methods using Keras.

1) U-Net training

Input of U-Net is RGB 3-channel images at 256×256 resolution. We used the cross-entropy (CE) loss as the objective function for training the U-Net model. The numbers of pixels of each class are unbalanced. Especially, the difference between crack and background is very large. Thus, we set 75:1:1 as the weights of crack class, boundary class, and background class, which is used to calculate cross entropy loss. Adaptive moment estimation (Adam) algorithm is used as the optimizer and the learning rate is 0.001. Minibatch size is 12. Then the number of epochs is fixed to 150 experimentally. Data augmentation was conducted for only training images, that includes random 90° rotations and random horizontal flips. These processes are applied randomly at a rate of 0.5, respectively.

2) Classification model training

Input of the classification model is RGB 3-chanel images at 48×48 resolution. The size of images is almost the same as the area used in the actual inspection. In the actual inspection, each 50 cm × 50 cm region in asphalt pavement surface is classified if cracks exist or not. Binary cross entropy (BCE) loss is used as the loss function in training process. Adam is used as the optimizer and the learning rate is 0.0001. Minibatch size is 32. The number of epochs is fixed to 100. Data augmentation was conducted for only training images in the same way as the U-Net training described above.

The block size of U-Net as the feature extractor is set 2, 3, and 4. The number of units in the fully connected layer is also set 32, 64, 128, 256, and 512. Total 15 combinations of parameters were evaluated and compared.

D. Evaluation

Fig. 4 describes the flow of evaluation of our proposed model. We used 100 images for training the CNN model and another 30 images for evaluation of classification performance of the trained model. These images were acquired on different routes on different days. The training images were divided to 2 sets of 50 images randomly. One of them were used to train the U-Net model as the feature extractor and the other of training images were used to train our classification CNN model. In both training of models, the rest 50 images in training dataset were used as validation data. We repeated 5 times procedure, to evaluate classification performance of our model with the average of precision, recall and F1-score on 5 trials.

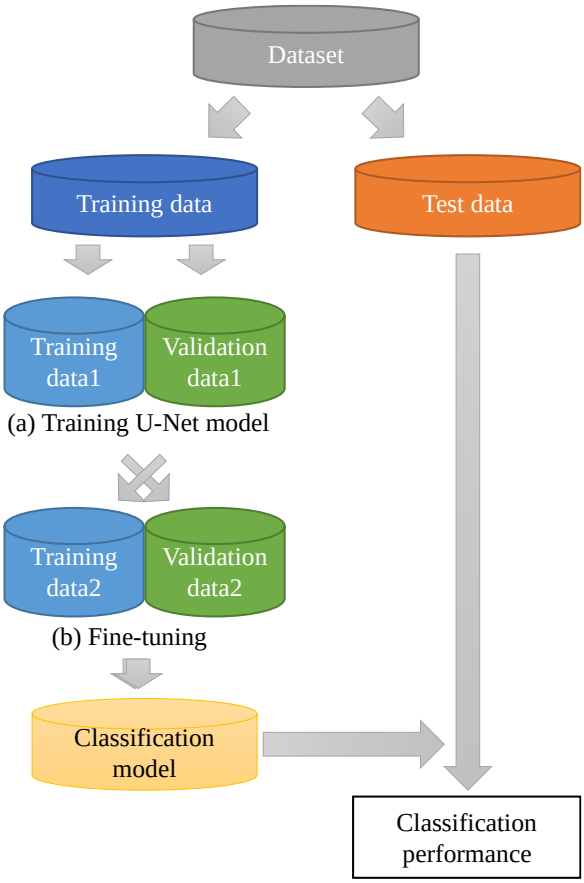


Figure 4. Flow of evaluation of our model.

E. Comparative Models

In our comparative evaluation, VGG11, ResNet18 and GoogLeNet are trained with Adam. Learning rate is set 0.001 and the number of epochs is set 100. The input images to each model were resized to the original input size of the model, respectively. Data augmentation was conducted for only training images in the same way as the training of our model described above.

IV. Result and Discussion

A. Classification Performance of Proposed Model

Firstly, we evaluated our proposed model, compare to the CNN model without U-Net pre-training whose architecture is same as our proposed model. The block size of the encoder of the U-Net as the feature extractor and the number of units in the fully connected layer are varied to compare the effect of them. Table I shows comparison of classification performance with varied parameters between our model and the comparative model, which was not pre-trained on U-Net. Overall, pre-training with the U-Net was shown to improve F1-score as the classification performance for all combinations of the block size and the number of units in the fully connected layer.

TABLE I. CLASSIFICATION PERFORMANCE OF OUR MODEL. THE BLOCK SIZE OF FEATURE EXTRACTOR AND THE NUMBERS OF UNITS IN FULLY CONNECTED LAYER ARE VARIED

Block size	# units of FC	Pre-training	Precision	Recall	F1-score
2	32		**0.546**	0.451	0.494
		✓	0.540	0.503	0.521
	64		0.541	0.473	0.504
		✓	0.502	0.541	0.521
	128		0.540	0.470	0.503
		✓	0.477	0.551	0.511
	256		0.530	0.469	0.498
		✓	0.517	**0.578**	**0.545**
	512		0.532	0.479	0.504
		✓	0.501	0.577	0.536
3	32		0.583	0.469	0.520
		✓	0.600	0.553	0.575
	64		0.524	0.462	0.491
		✓	**0.617**	0.551	0.582
	128		0.539	0.457	0.495
		✓	0.606	0.561	0.582
	256		0.553	0.475	0.511
		✓	0.609	0.559	0.583
	512		0.546	0.476	0.509
		✓	0.614	**0.566**	**0.589**
4	32		0.564	0.462	0.508
		✓	**0.698**	0.535	**0.605**
	64		0.555	0.462	0.504
		✓	0.665	0.535	0.593
	128		0.554	0.473	0.511
		✓	0.632	**0.557**	0.592
	256		0.535	0.468	0.499
		✓	0.617	0.576	0.596
	512		0.538	0.454	0.492
		✓	0.587	0.569	0.578

The highest F1-score of our model trained with pre-training is 0.605, whose block size of feature extractor and number of units in the fully connected layer of the model are 4 and 32, respectively. F1-score of our model increased from 0.508 to 0.605 by pre-training with U-net. It is also shown in Table II, to be compared to the other models.

B. Comparison with Transfer Learning Models

TABLE III shows the classification performance of VGG11, ResNet18, and GoogLeNet as comparative models, which are pre-trained with ImageNet or without transfer learning. GoogLeNet model using transfer learning obtained higher F1-score compared to that of the other models. In our experiment, while the amount of the training dataset is very small, transfer learning is effective for GoogLeNet, which is a very deep neural network model. However, transfer learning for VGG11 and ResNet18 was not effective in our experiment using the asphalt pavement dataset.

From Table II and Table III, comparison of our model and transfer leaning models indicates that pre-training of our model is more effective than that of GoogLeNet. For deep network such as GoogLeNet, transfer learning using large size of training data such as ImageNet is effective, especially in using the limited actual images for training process. However, the effect is limited in our experiment, it may be caused by the domain gap between natural images and asphalt pavement surface images. On the other hand, while our method required annotation for segmentation additionally, it is more effective to build CNN model with higher classification accuracy using the limited dataset.

TABLE II. CLASSIFICATION PERFORMANCE OF OUR MODEL. THE BLOCK SIZE OF FEATURE EXTRACTOR IS 4. THE NUMBER OF UNITS IN FULLY CONNECTED LAYER IS 32

Model and training method		Precision	Recall	F1-score
Our model	w/o Pre-training	0.564	0.462	0.508
	w/ Pre-training	**0.698**	**0.535**	**0.605**

TABLE III. CLASSIFICATION PERFORMANCE OF COMPARATIVE MODELS PRE-TRAINED USING IMAGENET AND WITHOUT PRE-TRAINING

Model and training method		Precision	Recall	F1-score
VGG11	w/o TL	0.491	0.490	0.490
	w/ TL	0.337	0.229	0.273
ResNet18	w/o TL	0.490	0.481	0.485
	w/ TL	0.331	0.231	0.272
GoogLeNet	w/o TL	0.470	**0.621**	0.535
	w/ TL	**0.688**	0.487	**0.570**

C. Limitations

- The architecture of our proposed model is selected experimentally, and it was not still optimized enough. Research of better architecture of CNN model for crack detection is important issues to overcome, to be expected to achieve higher performance.
- Our experiment was conducted using a limited very small size dataset. It is required to build our CNN model and evaluate the performance of the model using a larger amount of dataset, for applying to actual inspection. This is an issue for our future work.
- In our evaluation experiment, comparative experiments have not been conducted with other state-of-the-art methods. This is also an issue for our future work.

V. CONCLUSION

In this article, we have proposed a method to build a classification model based on the U-Net, which is the semantic segmentation model trained with a limited amount of the asphalt pavement surface dataset. We presented comparative evaluations with well-known models built by transfer learning using ImageNet. Experimental results show our model achieved to higher classification accuracy, compared to well-known models built by transfer learning using ImageNet. Our method is effective to build the CNN model with high classification accuracy using the limited dataset.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Y.F., Conceptualization, Methodology, Writing – original draft, Project administration; T.T., Software, Formal analysis, Validation; T.H., Software, Formal analysis, Validation; Y.H., Writing – review & editing, Supervision. All authors had approved the final version.

FUNDING

This work was supported by JSPS KAKENHI Grant Numbers JP18K04302 and JP22K04262.

ACKNOWLEDGMENT

The authors wish to thank Koji Shimada and Manabu Ichihara of Wesco Co. Ltd. for providing asphalt pavement surface dataset to conduct our research.

REFERENCES

[1] S. Chambon and J. M. Moliard, "Automatic road pavement assessment with image processing: Review and comparison," *International Journal of Geophysics*, art. no. 989354, 2011.
[2] C. Koch, K. Georgieva, V. Kasireddy, B. Akinci, and P. Fieguth, "A review on computer vision based defect detection and condition assessment of concrete and asphalt civil infrastructure," *Advanced Engineering Informatics*, vol. 29, no. 2, pp. 196–210, 2015.
[3] H. Zakeri, F. M. Nejad, and A. Fahimifar, "Image based techniques for crack detection, classification and quantification in asphalt pavement: A review," *Arch Computat Methods Eng*, vol. 24, no. 4, pp. 935–977, 2017.
[4] Y. Hou, Q. Li, C. Zhang, G. Lu, Z. Ye, Y. Chen, L. Wang, and D. Cao, "The state-of-the art review on applications of intrusive sensing, image processing techniques, and machine learning methods in pavement monitoring and analysis," *Engineering*, vol. 7, pp. 845–856, 2021.
[5] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Proc. of the 27th NeurIPS*, 2014, vol. 2, pp. 3320–3328.
[6] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, pp. 211–252, 2015.
[7] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. of MICCAI 2015*, pp. 234–241.
[8] P. J. Chun, T. Yamane, and Y. Tsuzuki, "Automatic detection of cracks in asphalt pavement using deep learning to overcome weakness in images and GIS visualization," *Applied Sciences*, vol. 11, no. 3, p. 892, 2021, doi: https://doi.org/10.3390/app11030892
[9] F. Liu, J. Liu, and L. Wang, "Deep learning and infrared thermography for asphalt pavement crack severity classification," *Automation in Construction*, vol. 140, 2022.
[10] S. Minaee, Y. Boyko, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE TPAMI*, vol. 44, no. 7, pp. 3523–3542, 2022.
[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. on ICLR*, 2015.
[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. of the IEEE Conf. on CVPR*, 2016, pp. 770–778.
[13] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. of the IEEE Conf. on CVPR*, 2015, pp. 1–9.