Asphalt Pavement Crack Detection Based on Convolutional Neural Network and Infrared Thermography

Fangyu Liu, Jian Liu, and Linbing Wang

***Abstract*— Two issues exist in the convolutional neural network (CNN) used for asphalt pavement crack detection: balance between accuracy and complexity, and indistinct edges of cracks and asphalt pavement surface. Infrared thermography (IRT) could use the temperature difference between cracks and pavement surface to better distinguish cracks. This work aims to propose a robust crack detection method based on CNN and IRT. An open benchmark dataset was built for crack detection based on three types of images, including visible images, infrared images, and the fusion of visible and infrared images (in short, fusion image). The dataset also considers different conditions and periods, including single, multi, thin, and thick cracks; clean, rough, light, and dark backgrounds, and three periods in a day. Seven CNN segmentation models are trained and evaluated on this dataset. To keep a balance between accuracy and complexity, evaluation metrics (accuracy, and computational and model complexity) are used to have an overall evaluation of models rather than only the accuracy. The results show that the accuracy and predictions of the visible image and fusion image are almost identical for all models, which are much better than that of the infrared image. When the background is rough or cracks are similar to the background, the fusion image is a better choice for crack detection. Compared with the visible image, all segmentation models have a more stable performance for the fusion image. Among segmentation models, Feature Pyramid Networks (FPN) could be the best model because of its high accuracy and low complexity.**

***Index Terms*— Crack detection, convolutional neural network, infrared thermography, asphalt pavement.**

I. Introduction

HOW to more efficiently and accurately detect pavement cracks? Being subjected to heavy and repetitive loading, and undesirable environmental impact (e.g. freeze-thaw), pavements need periodic inspections, assessments, maintenances, and rehabilitation to ensure their safety and prolong their service life [1], [2]. Crack is one of the crucial distresses for pavement because crack is the common symptom of pavement [3]. Different from other civil infrastructures, pavement crack types could be mainly separated into three categories: transverse crack, longitudinal crack, and alligator crack, which may be attributed to different causes [4], [5]. Pavement cracks could reduce the load-spreading and water-resistant capacity of pavement as well as accelerate the degradation process of pavement surface [6]. Pavement crack detection strongly depends on manual inspection and contact- or embedded-sensor-based methods [3], [7]. However, manual inspection is susceptible, unreliable, and time-consuming while it is dangerous in some cases [7]. The application of sensor-based methods is also limited due to environmental disadvantages and vulnerability to temperature and humidity changes [3]. Therefore, a safe, reliable, accurate, non-contact, and efficient method is needed for pavement crack detection.

The convolutional neural network (CNN) could be an ideal technique for crack detection. With the rapid development of deep learning, CNN has become a powerful and efficient tool for computer vision tasks, such as image recognition, semantic segmentation, and object detection. Several studies have illustrated the validity and usefulness of CNN in detecting pavement cracks [2], [8]–[11]. All these works evaluated their methods and compared their methods with previous methods based on the accuracy criterion (e.g. precision and F_1 score). However, only the accuracy could not fully illustrate the performance of CNN models. The model and computational complexity also play an important role in evaluating CNN [12], [13]. Although some time-consuming CNN models achieve state-of-the-art performance, they could be unaffordable or unnecessary for engineering applications [12]. These problems highlight the importance of balance between accuracy and complexity and urge to find an accurate and efficient CNN model for detecting pavement cracks.

Another issue is the indistinct edges of cracks and asphalt pavement surface [14], [15]. Asphalt pavement is bonded granular material; randomly distributed particles make the edges of cracks and pavement surface indistinct [14] (Fig. 1a). Therefore it may cause misidentification of many noises in the background as crack fragments [15]. Infrared thermography (IRT), as a contactless, non-invasive, and non-destructive testing method, has been successfully applied in civil engineering, such as assessment of HMA paving and compaction [16], debonding detection in asphalt pavements [17], and defect identification in composite materials [18]. Different from visible images that only provide gray-level information, IRT could generate gray-level information and temperature infor-

Manuscript received 4 August 2021; revised 21 October 2021; accepted 31 December 2021. Date of publication 19 January 2022; date of current version 7 November 2022. This work was sponsored by a grant from the Center for Integrated Asset Management for Multimodal Transportation Infrastructure Systems (CIAMTIS), a US Department of Transportation University Transportation Center, under federal grant number 69A3551847103. The Associate Editor for this article was X. Luo. *(Corresponding author: Linbing Wang.)*

The authors are with the Department of Civil and Environmental Engineering, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061 USA (e-mail: lfangyu@vt.edu; jian19@vt.edu; wangl@vt.edu).

Digital Object Identifier 10.1109/TITS.2022.3142393

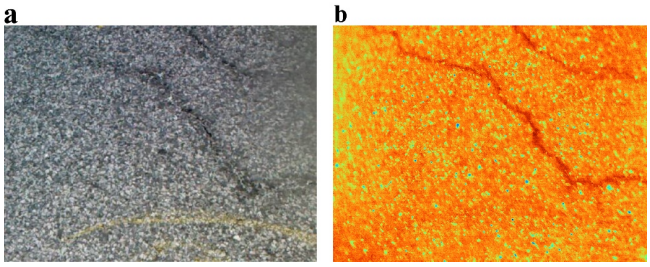

Fig. 1. Examples of (a) the visible image and (b) the infrared image.

mation [14], [19]. The significant temperature difference between cracks and pavement surfaces [14], [16], [20] could better distinguish cracks in infrared images than visible images (Fig. 1), thereby improving the detection accuracy.

Therefore, this work aims to propose a robust crack detection method based on CNN and infrared thermography. The following summarizes the primary contribution of this work:

- Build an open benchmark dataset for crack detection based on three types of images, including visible images, infrared images, and the fusion of visible and infrared images (in short, fusion images).
- The dataset also considers different conditions and periods, including single and multi cracks, thin and thick cracks, clean and rough backgrounds, light and dark backgrounds, and three periods in a day.
- Five classical CNN segmentation models and two UNet-based models are trained and evaluated on the aforementioned dataset to make a benchmark.
- Evaluation metrics (accuracy, computational complexity, and model complexity) are used to have an overall evaluation of CNN segmentation models rather than the only accuracy metrics.

II. Methods

A. Infrared Camera

The infrared camera is FLUKE TiX580 and has two lenses, including a visible light camera lens and an infrared camera lens (Fig. 2a). The two lenses enable this infrared camera to take the visible image and infrared image simultaneously. The image size is 640×480. In addition to only visible images and infrared images, the fusion of visible and infrared images could be obtained by using the IR-Fusion^{TM} technology to adjust the blending level of visible images and infrared images. The display screen in the back of the infrared camera (Fig. 2b) could show the image to help to adjust the camera angle and camera range. In addition to visualization, it could also edit emissivity, enable color alarms and makers, and adjust the blending level of visible and infrared images. After storing images on computers, the software (SmartView) could also adjust the level of visible and infrared light (Fig. 2c).

B. Segmentation Models

Crack detection could be divided into two main categories: object detection and segmentation. Object detection needs to distinguish objects in an image or video and segmentation needs to label each pixel to certain classes. Specifically for crack detection, object detection often points out the location and types of cracks. Different from object detection, segmentation would label each pixel as 'crack' or 'non-crack' and its outputs are usually binary images, while white ('1') often represents 'crack' and black ('0') means 'non-crack'. This method could be used for qualitative and quantitative analysis, such as the level and direction of cracks. The work presented in this paper would focus on segmentation in terms of crack detection. As an approach for automatically detecting important features without human supervision, CNN has been widely applied in image segmentation. Common CNN models for this application include Pyramid Scene Parsing Network (PSPNet) [21], Feature Pyramid Networks (FPN) [22], Fully Convolutional Networks (FCN) [23], DeepLab [24], and UNet [25]. PSPNet was a multi-scale network and combined up-sampling and concatenation to capture both local and global context information [21]. FPN was mainly proposed for object detection and then was also applied in image segmentation, while it consisted of a bottom-up pathway, a top-down pathway, and lateral connections [22]. FCN, as one of the first CNN models for image segmentation, replaced the fully connected layers of typical CNN architectures with fully convolutional layers and combined semantic information and appearance information by skip connections [23]. DeepLab used the atrous spatial pyramid pooling (ASPP) and combined the cascade and parallel modules of dilated convolution [24]. UNet was initially proposed for the segmentation of biological microscope images and had two parts, encoder (down-sampling) and decoder (up-sampling) [25]. UNet-based models (UNet-ResNet101 and UNet-VGG19) were built by replacing the encoder of the UNet with typical CNN architectures. The encoders of FCN, FPN, and PSPNet are VGG-19 [26] in this work, while the encoder of DeepLabv3 is ResNet-101 [27]. Models are established based on [28]–[30].

C. Evaluation Metrics

To better evaluate the performance of segmentation models and keep a balance between accuracy and complexity, three evaluation metrics are used, including accuracy, computational complexity, and model complexity.

1) Accuracy: Many criteria have been proposed to evaluate the accuracy of segmentation models and pixel accuracy, mean pixel accuracy, and mean intersection over union are the most popular accuracy metrics used to measure the performance of per-pixel labeling methods on image segmentation [13]. Pixel accuracy (PA) calculates the ratio between the number of correctly classified pixels and their total number and mean pixel accuracy (MPA) is an improved PA that considers the per-class basis [13]. Mean intersection over union (MIoU) calculates the ratio between the intersection and the union of two sets (the ground truth and predicted segmentation) and it is the most used metrics because of its representativeness and simplicity [13]. In addition, the F_1 score is the measure of accuracy in binary classification. Suppose there are a total of $k + 1$ classes (2 classes for crack detection, 'crack' and 'non-crack') and p_{ij} refers to the number of pixels of class

Fig. 2. Overview of the infrared camera and software. (a) The front of the infrared camera (including the visible light camera lens and infrared camera lens), (b) the back of the infrared camera, (c) the software used to process images (including temperature information and blending level of images).

i inferred to belong to class j. This means p_{ii}, p_{ij}, and p_{ji} represent the number of true positives (TP), false positives (FP), and false negatives (FN), respectively.

- Pixel Accuracy (PA):

$$PA = \frac{\sum_{i=0}^{k} p_{ii}}{\sum_{i=0}^{k}\sum_{j=0}^{k} p_{ij}} \tag{1}$$

- Mean Pixel Accuracy (MPA):

$$MPA = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k} p_{ij}} \tag{2}$$

- Mean Intersection over Union (MIoU):

$$MIoU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}} \tag{3}$$

- F_1 score:

$$F_1 score = 2\times\frac{Precision \cdot Recall}{Precision + Recall} = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{4}$$

2) Computational and Model Complexity: The computational complexity of CNN reflects the speed or runtime. It is a very valuable metric as most CNN models must meet strict requirements on how much time can be spent on the inference pass [13]. However, the exact timings might be meaningless because it strongly relies on the hardware. Depending on the architecture of CNN, the floating-point operations (FLOPs) could represent the execution time or computational complexity of CNN [31]. Therefore, FLOPs are used to evaluate the computational complexity of segmentation models in this work.

Another important evaluation metric is the model complexity. This metric heavily depends on the number of the model's parameters. More parameters mean the model is more complex and needs more training data to avoid overfitting problems. As almost all deep learning methods need Graphics Processing Units (GPUs) to speed up, the GPU memory proposes the requirements on the model complexity. Generally speaking, the higher the model complexity, the more GPU memory is required. This might be a challenge to a research group. The FLOPs and the number of parameters are computed based on [32].

D. Hyperparameters of Models

The segmentation models were trained and evaluated on PyTorch [33]. The inputs are RGB images (size: $640\times480\times3$) and the outputs or ground truths are binary images (size: $640\times480\times1$). The loss function is the Binary Cross-Entropy (BCE) loss and the optimization algorithm is Adam [34]. Other hyperparameters include the epoch (100), learning rate (0.0001), and batch size (1, mini-batch). For all epochs, the accuracy metrics of the model were immediately evaluated after training. The model was saved every five epochs.

III. Dataset

There is rarely an open benchmark dataset for crack detection based on IRT, although few previous studies provided small datasets for crack detection. Therefore, an open benchmark dataset for IRT-based crack detection was established in this work.[1] There are four types of images in this dataset, including the visible image, infrared image, fusion image, and ground truth. The visible image and infrared image are fully visible and infrared, respectively. The fusion image is the combination of visible and infrared images achieved by IR-Fusion^{TM} technology, which is approximately 50% and 50%, respectively. According to the previous literature [7], [35], [36], the ground truth in the dataset is manually labeled at the pixel level by using Photoshop software.

As the temperature difference between cracks and pavement surfaces enables IRT to distinguish cracks, the temperature change in the whole day might influence this temperature difference. Therefore, data (images) were acquired at three periods, including morning (8:00 am), noon (12:00 pm), and dusk (5:00 pm). The highest pavement surface temperature is almost consistent with the daily maximal temperature, while the dusk temperature is a little higher than that of the morning. Fig. 3 shows examples of the visible image, infrared image, and fusion image at the same location in three periods. Although these visible images are greatly similar and even identical, their infrared images and fusion images are significantly different. Firstly, their temperatures are different. Images at noon (Fig. 3b) have the highest temperature, ranging from 17.1 to 31.0 °C, while images at dusk (Fig. 3c) have the second-highest temperature (5.0 to 15.9 °C) and images

[1] Available at: https://github.com/lfangyu09/IR-Crack-detection

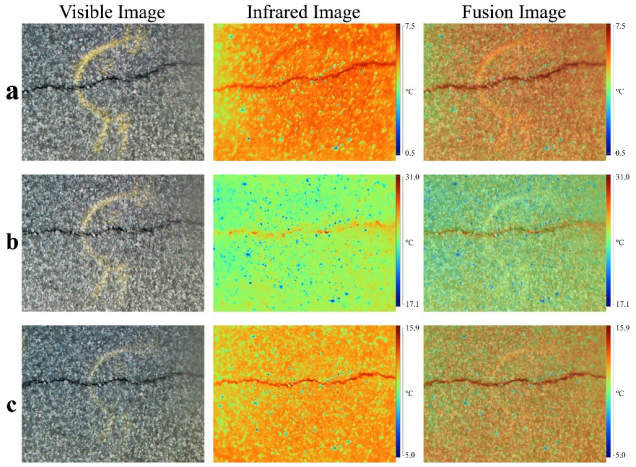

Fig. 3. Examples of the visible image, infrared image, and fusion image at the same location in three periods. (a) Morning, (b) noon, and (c) dusk.

in the morning (Fig. 3a) have the lowest temperature (0.5 to 7.5 °C). This means the same color in these images represents different temperatures. Secondly, the crack area has the highest temperature but the background has different patterns. The image background at noon approximately has the middle temperature of its temperature range. The image background temperature at dusk is higher than the middle temperature and so does the image background temperature in the morning, while its color is much closer to red (much higher than the middle temperature). Thirdly, the distinction of cracks is different. Some crack areas in images at noon have a similar temperature with the background, causing some misidentification of cracks. The distinction of cracks in images at dusk and in the morning is more clear and obvious. In addition, shadows caused by guardrails and trees might affect both visible images and infrared images. To eliminate such environmental factors, images were only taken from the pavement without guardrails and trees.

Fig. 4 shows examples of the visible image, infrared image, fusion image, and ground truth in the dataset. There are eight types of conditions, including single and multi cracks, thin and thick cracks, clean and rough backgrounds, and light and dark backgrounds. These different images are used to evaluate the performance of segmentation models under different conditions. It can be observed that in the infrared image or fusion image, it is easy to distinguish cracks when the width of cracks is small (Fig. 4a-c, g, and h). When the width of cracks increases, the boundary of cracks and the background is not clear and there is a buffer zone, namely color gradient areas, between these two areas (Fig. 4d and e). Another phenomenon that needs to be figured out is that the wet area generally has a lower temperature (Fig. 4d and f).

Table I illustrates the summary of this dataset. There are a total of 448 images for each type (4 types, including the visible image, infrared image, fusion image, and ground truth). 186 images were taken in the morning, 142 at noon, and 120 at dusk. To train and evaluate the segmentation models, the whole dataset is separated into two subsets, the training set, and the test set. The training set has 382 images and the test set has 66 images. In addition, the percentages of crack pixels and non-crack pixels are also summarized in Table I. The crack area only occupies a small part (less than 4%) of the whole image.

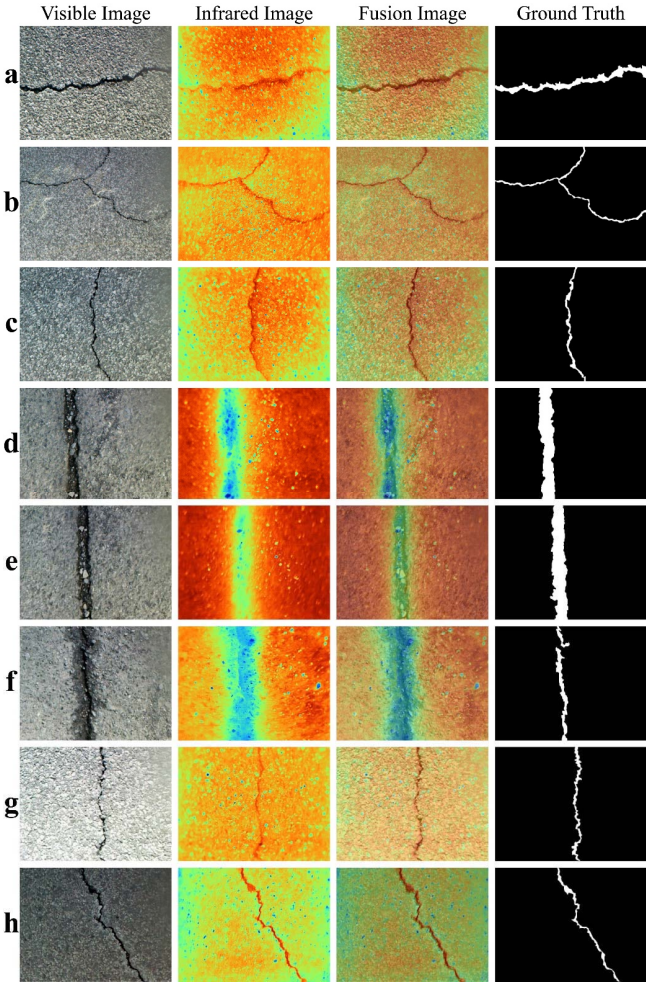

Fig. 4. Examples of the visible image, infrared image, fusion image, and ground truth in the dataset. (a) Single crack, (b) multi cracks, (c) thin crack, (d) thick crack, (e) clean background, (f) rough background, (g) light background, and (h) dark background.

TABLE I
SUMMARY OF THE DATASET

Dataset	Number	Crack Pixel (%)	Non-crack Pixel (%)
Morning	186	3.85	96.15
Noon	142	3.97	96.03
Dusk	120	3.20	96.80
Train	382	3.86	96.14
Test	66	2.88	97.12
Total	448	3.71	96.29

IV. Results and Discussion

A. Loss

Fig. 5 shows the loss curves of the visible image, infrared image, and fusion image. These three types of images have different patterns for the loss curves. For the visible image (Fig. 5a), the loss of almost all segmentation models tends to converge with the increasing epochs but the loss curve

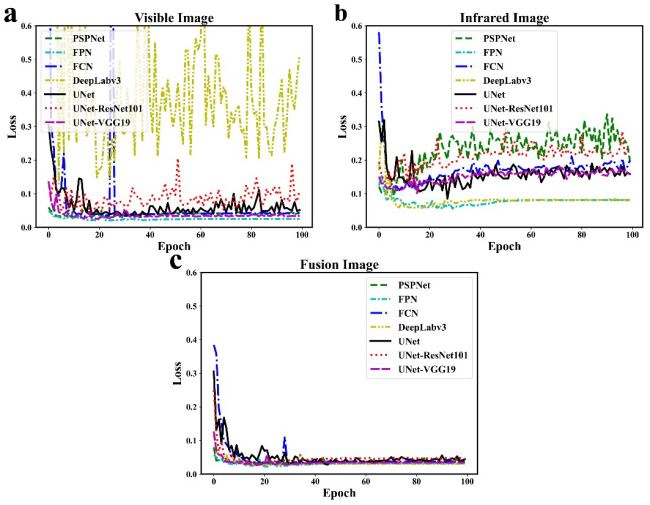

Fig. 5. Loss curves of the test set. (a) Visible images, (b) infrared images, (c) fusion images.

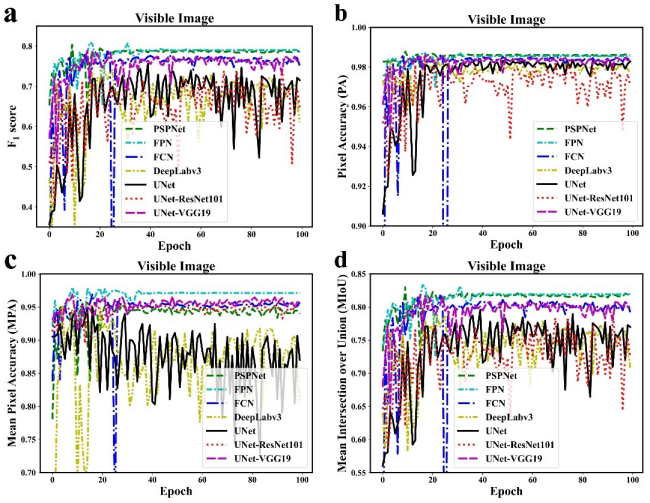

Fig. 6. Accuracy metrics of the visible image. (a) F_1 score, (b) PA, (c) MPA, and (d) MIoU.

of DeepLabv3 experiences huge fluctuations, while UNet-ResNet101 has a slight fluctuation. Different from the visible image, the loss curve of DeepLabv3 for the infrared image (Fig. 5b) tends to be stable at a low value as the epoch increases and so does the loss curve of FPN. Other segmentation models face the overfitting problem to some degree. Their loss curves firstly decrease and then gradually increase, especially PSPNet and UNet-ResNet101. To eliminate the effect of the overfitting problem, the final results of models were selected as the best results during all epochs rather than the results of the final epoch. The loss curves for the fusion image (Fig. 5c) have a different pattern. All segmentation models tend to converge when the epoch increases and their loss curves are almost overlapped at a very small value after Epoch 40. From the perspective of the loss curve, the fusion image is a better choice as the loss curves of all segmentation models could converge at a small value.

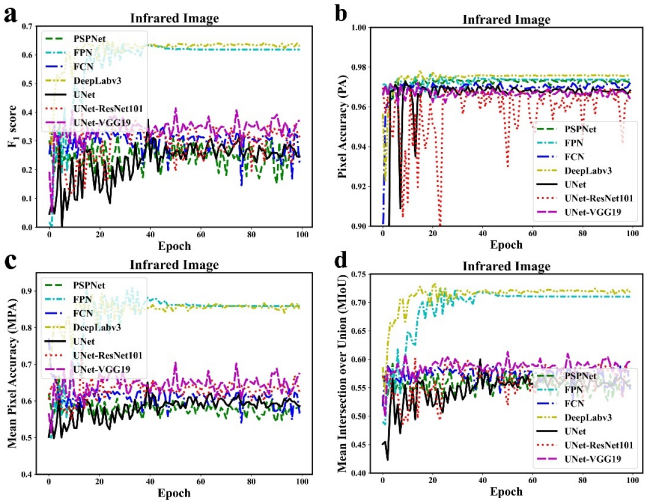

Fig. 7. Accuracy metrics of the infrared image. (a) F_1 score, (b) PA, (c) MPA, and (d) MIoU.

B. Evaluation Metrics

1) Accuracy: The outputs (predicted probability maps) of segmentation models range from 0 to 1. To get the binary image and compute the accuracy, it needs to binarize these outputs to 0 or 1 with thresholds. 0.5 was selected as the threshold as it is often used in the previous literature [36] and is also simple.

Fig. 6 shows the curves of accuracy metrics of the visible image, including F_1 score, PA, MPA, and MIoU. For the pixel accuracy (Fig. 6b), these segmentation models tend to be stable and reach a high value after Epoch 40, except for UNet-ResNet101. In comparison, for other accuracy metrics UNet, DeepLabv3, and UNet-ResNet101 have experienced huge fluctuations while UNet-ResNet101 has a good performance in MPA (Fig. 6c). The other segmentation models tended to stabilize in all accuracy metrics as the epoch increases. Table II summarizes the final performance of segmentation models on the visible image. FPN obtains the highest value in three accuracy metrics (F_1 score, MPA, and MIoU) and it is 0.001 smaller than PSPNet in PA. In comparison, UNet-ResNet101 has the lowest value in almost all accuracy metrics except for MPA. According to the performance during the training and final results, FPN is the best segmentation model for the visible image among these segmentation models.

The results of the infrared image show a different pattern. Fig. 7 shows the curves of accuracy metrics of infrared images. The values of FPN and DeepLabv3 are much higher (significant gap) than that of other segmentation models in F_1 score, MPA, and MIoU, while all segmentation models have similar performance in PA (Fig. 7b) except for UNet-ResNet101. There is a little difference between other segmentation models in F_1 score, MPA, and MIoU. Table III illustrates the final results of the segmentation models on the infrared image. The value of all accuracy metrics for the infrared image is much lower than that of the visible image. DeepLabv3 achieves the highest value in all accuracy metrics for the infrared image although it has the second-lowest value in all accuracy metrics for the visible image. All accuracy metrics of FPN are slightly lower than that of DeepLabv3. These two models have a far better performance than the other segmentation models.

TABLE II
PERFORMANCE OF DIFFERENT MODELS ON THE VISIBLE IMAGE

Model	Accuracy Metrics				Parameters (M)	FLOPs ($1x10^9$)
	F_1	PA	MPA	MIoU		
PSPNet [21]	0.804	0.988	0.930	0.830	20.82	111
FPN [22]	0.809	0.987	0.963	0.833	22.11	131
FCN [23]	0.788	0.986	0.949	0.818	23.93	199
DeepLabv3 [24]	0.751	0.984	0.925	0.792	58.63	283
UNet [25]	0.756	0.984	0.923	0.796	17.27	188
UNet-ResNet101 [28]	0.747	0.983	0.933	0.790	51.51	73
UNet-VGG19 [28]	0.789	0.986	0.947	0.818	29.06	142

TABLE III
PERFORMANCE OF DIFFERENT MODELS ON THE INFRARED IMAGE

Model	Accuracy Metrics				Parameters (M)	FLOPs ($1x10^9$)
	F_1	PA	MPA	MIoU		
PSPNet [21]	0.362	0.964	0.668	0.592	20.82	111
FPN [22]	0.642	0.975	0.880	0.724	22.11	131
FCN [23]	0.347	0.969	0.638	0.589	23.93	199
DeepLabv3 [24]	0.658	0.976	0.893	0.733	58.63	283
UNet [25]	0.375	0.970	0.650	0.600	17.27	188
UNet-ResNet101 [28]	0.400	0.956	0.740	0.603	51.51	73
UNet-VGG19 [28]	0.415	0.965	0.704	0.613	29.06	142

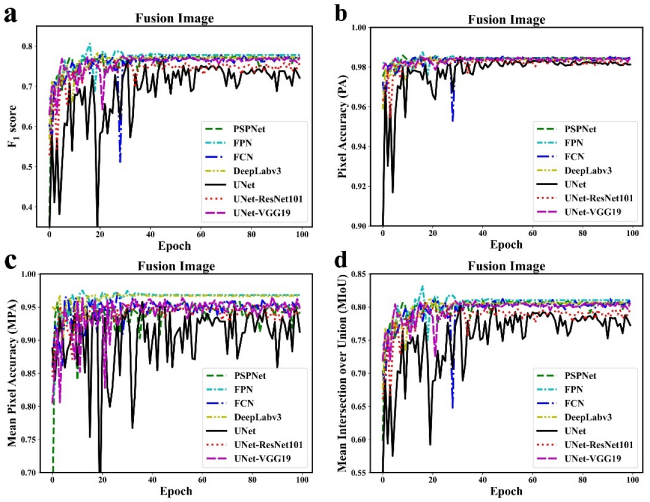

Fig. 8. Accuracy metrics of the fusion image. (a) F_1 score, (b) PA, (c) MPA, and (d) MIoU.

The results of the fusion image are similar to that of the visible images. Fig. 8 shows the curves of accuracy metrics for the fusion image. Although UNet reveals some fluctuations in accuracy metrics, the other segmentation models tend to stabilize at a high value. Compared with the visible image, all segmentation models have a more stable performance in the fusion image. Table IV reveals the final results of segmentation models on the fusion image. Similar to the visible image, FPN has the highest value in three accuracy metrics (F_1 score, PA, and MIoU). The high value of MPA has been observed in DeepLabv3 whose values of other accuracy metrics are a little lower than that of FPN. This is different from the results of the visible image.

To better present the experimental results, Fig. 9 summarizes the results of the accuracy metrics of segmentation models for three types of images. In general, the accuracy metrics of the visible image and fusion image are similar and even identical for all segmentation models, which are much better than that of the infrared image. FPN has a satisfying performance in all three types of images. And DeepLabv3 has a good performance in the infrared image and fusion image while compared with other segmentation models, its performance in the visible image is relatively poor but acceptable. The difference in the same accuracy metrics was caused by two main aspects: (1) the type of datasets. For the visible image, it could work as the baseline in this work as it has been always used in all segmentation tasks until now. For the infrared image, when the width of cracks is small, the boundaries between cracks and the background are very clear (Fig. 4a-c, g, and h). But the boundaries fade out and the buffer zones (color gradient areas) appear as cracks become wider (Fig. 4d and e). These phenomena increase the difficulty of crack detection and make the same segmentation models behave differently in the visible image and infrared image. For the fusion image, as it integrates 50% of the visible image and 50% of the infrared image, the boundaries keep clear no matter whether cracks are thin or wide. This enables segmentation models to have similar performances in the fusion image with the visible image. In addition, with the help of the gray-level information and temperature information [14], [19], the fusion image makes segmentation models more stable (Fig. 5 and Fig. 8). (2) the type of segmentation models. For CNN, shallower feature maps are of lower-level semantics but more accurately localized, while deeper feature maps are semantically stronger but spatially coarser [22], [37]. FPN merges

TABLE IV
PERFORMANCE OF DIFFERENT MODELS ON THE FUSION IMAGE

Model	Accuracy Metrics				Parameters (M)	FLOPs ($1x10^9$)
	F_1	PA	MPA	MIoU		
PSPNet [21]	0.780	0.986	0.938	0.812	20.82	111
FPN [22]	0.806	0.987	0.948	0.831	22.11	131
FCN [23]	0.780	0.985	0.952	0.812	23.93	199
DeepLabv3 [24]	0.782	0.985	0.959	0.814	58.63	283
UNet [25]	0.768	0.984	0.950	0.804	17.27	188
UNet-ResNet101 [28]	0.771	0.985	0.942	0.806	51.51	73
UNet-VGG19 [28]	0.776	0.985	0.949	0.809	29.06	142

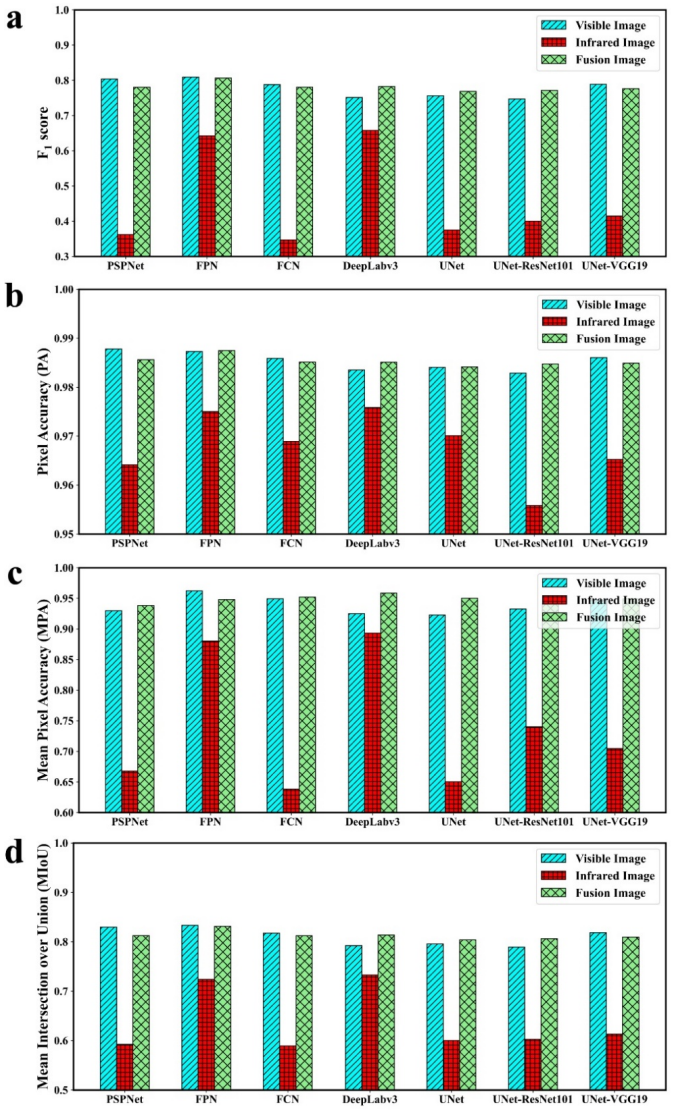

Fig. 9. Accuracy metrics of segmentation models for three types of images. (a) F_1 score, (b) PA, (c) MPA, and (d) MIoU.

these two types of features to maximize their corresponding advantages, namely, the high resolution from shallower feature maps and the strong semantics from deeper feature maps [22], [37]. In addition, different from the traditional top-down architecture (e.g. UNet) which makes predictions on the finest level, FPN regards the top-down architecture as a feature pyramid and makes predictions independently at all levels [22]. Therefore, FPN shows a prominent performance in all three types of images. For DeepLab v3, it used the cascaded module to double the atrous rates while it also applies the atrous spatial pyramid pooling module augmented with image-level features for detecting features [38]. These enable DeepLab v3 to maintain a good performance when the type of the input image changes.

In addition to the types of segmentation models and images, these four accuracy metrics also influence evaluating the performance of segmentation models on crack detection. The PA of all images is very high (>0.95) and much larger than the other accuracy metrics. The three types of images do not have a huge difference in PA, and the largest gap among all segmentation models is less than 0.04. In comparison, for the other accuracy metrics, there is an obvious and big difference between these three types of images for all segmentation models and the largest difference is up to 0.44 in the F_1 score. This is because the pixel accuracy only computes the correctly classified pixels without considering the effect of the class ('crack' and 'non-crack'), while the other accuracy metrics (F_1 score, MPA, and MIoU) take the class into account to get the final value. The ground truth of crack detection (Fig. 4) is the class imbalance and the crack pixel only occupies a very little part (less than 4%) of the whole image (Table I). The high value (>0.95) of PA and low difference (<0.04) between the three types of images are mainly contributed to the correct classification of the background. After considering the effect of the class, the value of MPA decreases, and the difference between the three types of images increases. The F_1 score and MIoU have already considered the effect of the class in their formula. The effect of the class would distinguish the difference between the three types of images. However, crack detection mainly focuses on the crack area and this needs more emphasis on the class of 'crack' rather than all classes. In this case, the visible image and fusion image are better choices for crack detection than the infrared image.

2) Computational and Model Complexity: Table II-IV show the computational and model complexity. As they are the internal characteristics of models, they are not affected by the types of images and they are the same for the three types of images. As mentioned above, the computational complexity is evaluated by FLOPs and the model complexity is determined by the number of parameters. DeepLabv3 has the highest value in the number of parameters and FLOPs.

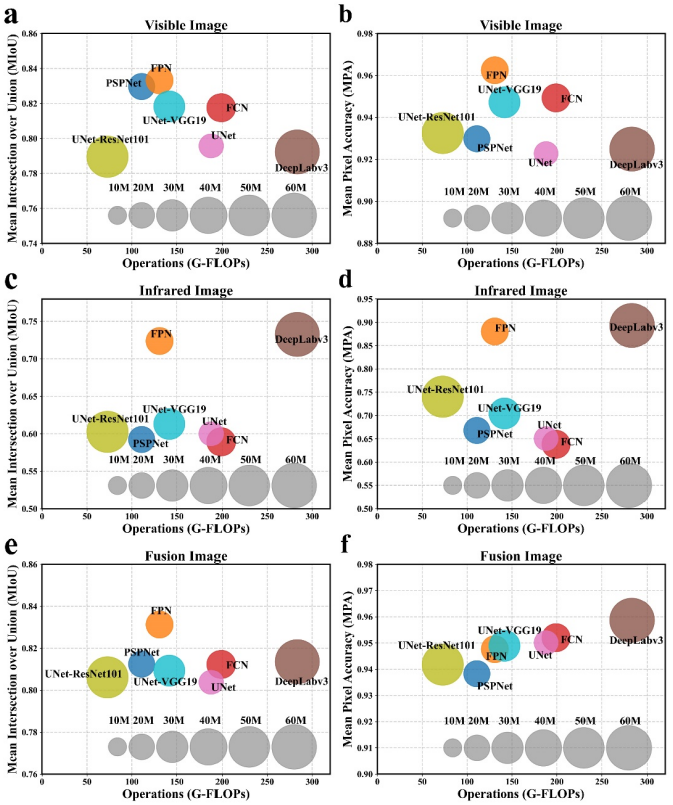

Fig. 10. Ball chart reporting the accuracy metrics (MIoU and MPA) vs. computational complexity (floating-point operations (FLOPs), G-FLOPs $= 1 \times 10^9$ FLOPs) with the same threshold. The size of each ball corresponds to the number of parameters for each model. (a) MIoU and (b) MPA for the visible image, (c) MIoU and (d) MPA for the infrared image, (e) MIoU and (f) MPA for the fusion image.

This means DeepLabv3 spends more execution time and also needs more training data to reduce the risk of overfitting. UNet has the lowest number of parameters, which is one-third of the value of DeepLabv3. The lowest FLOPs are observed in UNet-ResNet101, which is almost one-quarter of the value of DeepLabv3, but the number of parameters for UNet-ResNet101 is approximately three times that of UNet. The FLOPs of UNet are about 2.5 times bigger than that of UNet-ResNet101. FPN has the third-lowest value in these two metrics. It could be found that there is no significant pattern between the number of parameters and FLOPs.

3) Accuracy vs. Computational and Model Complexity: To comprehensively evaluate the performance of segmentation models and the types of images, accuracy metrics, computational complexity, and model complexity are combined to make an analysis, and the results are shown in Fig. 10. Considering the effect of the class imbalance, MIoU and MPA are selected to represent the accuracy. The best model in the three types of images is different while FPN could be ranked the top two in almost all accuracy metrics with the low FLOPs and the number of parameters. With the highest FLOPs (x coordinates) and the number of parameters (the size of the ball), DeepLabv3 could be the top two in the infrared image and fusion image but it is almost the worst model for the visible image. UNet-VGG19 could also be a good alternative for detecting cracks in the three types of images.

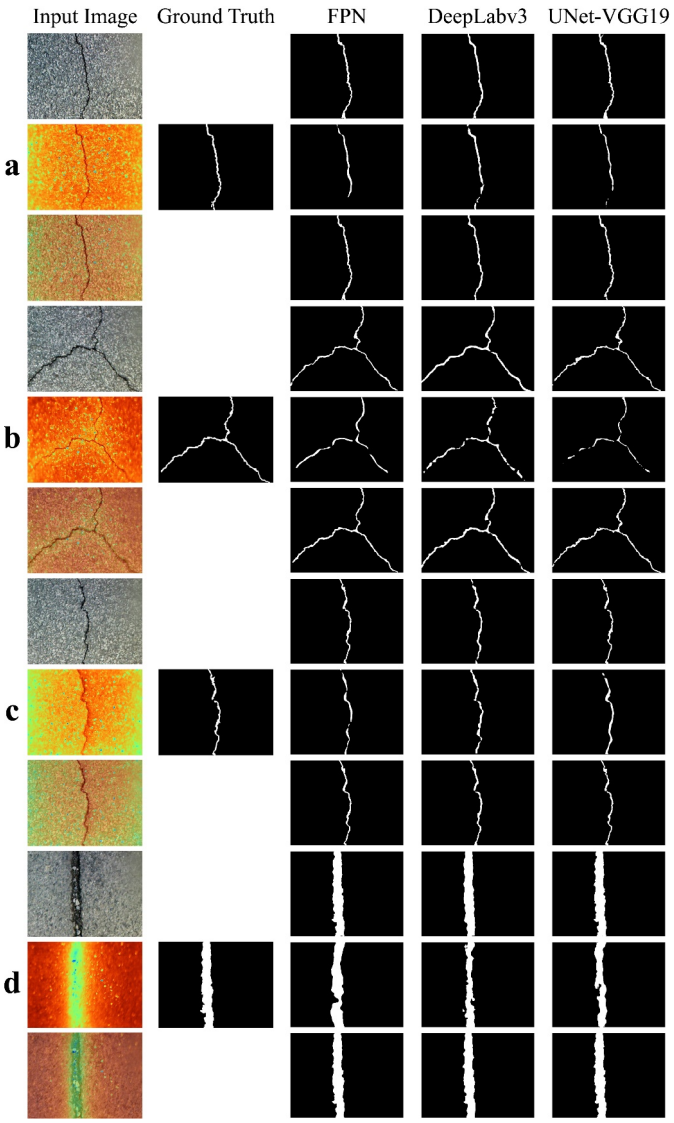

Fig. 11. Predictions of three segmentation models for different cracks based on three types of images. (a) Single crack, (b) multi cracks, (c) thin cracks, and (d) thick cracks.

C. Prediction

In addition to evaluation metrics, prediction is also important to evaluate the performance of segmentation models as it could provide an intuitive assessment. As mentioned above, various conditions have been considered to build the dataset. The prediction under different conditions is used to further evaluate the performance of segmentation models. Based on evaluation metrics, three segmentation models (FPN, DeepLabv3, and UNet-VGG19) are selected to show their predictions on different conditions.

Fig. 11 shows the predictions of three segmentation models for different cracks based on three types of images. In short, the predictions of the visible image and fusion image are almost identical for all segmentation models, which are much better than the infrared image, especially for multi and thick cracks. As for the performance of the three segmentation models, there is no significant difference in terms of the visible image and fusion image and their predictions are

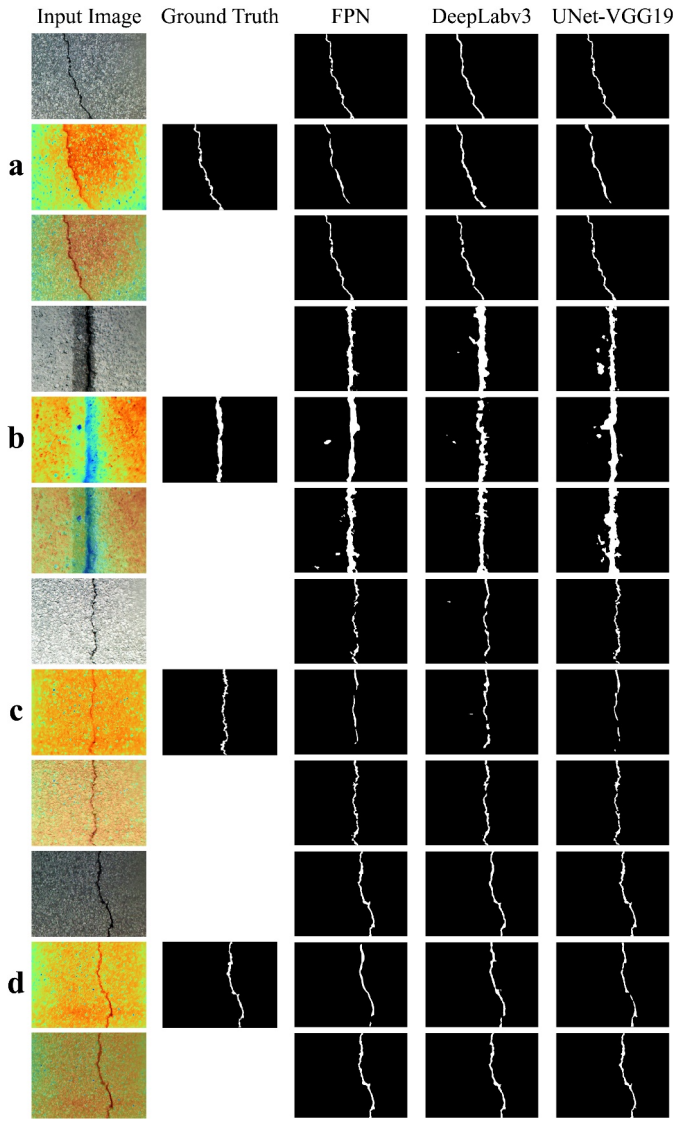

Fig. 12. Predictions of three segmentation models for different backgrounds based on three types of images. (a) Clean background, (b) rough background, (c) light background, and (d) dark background.

almost identical to the ground truth. DeepLabv3 has a much better prediction of the infrared image than the other two segmentation models.

The dataset also includes various backgrounds. Fig. 12 shows the predictions of three segmentation models for different backgrounds based on three types of images. The predictions of the visible image and fusion image are still better than that of the infrared image for all segmentation models, while the visible image has a similar prediction with that of the fusion image. However, it is worth noting that when the background is rough, such as with water, the fusion image might be a better choice. The color of the pavement wetted by water is very close to that of cracks (Fig. 4f and Fig. 12b), thereby increasing the difficulty of detecting cracks based on the visible image. This is why the predictions of the three segmentation models are unsatisfactory. In comparison, water decreases the temperature and turns the wet areas blue in the infrared image and the fusion image (Fig. 4f and Fig. 12b).

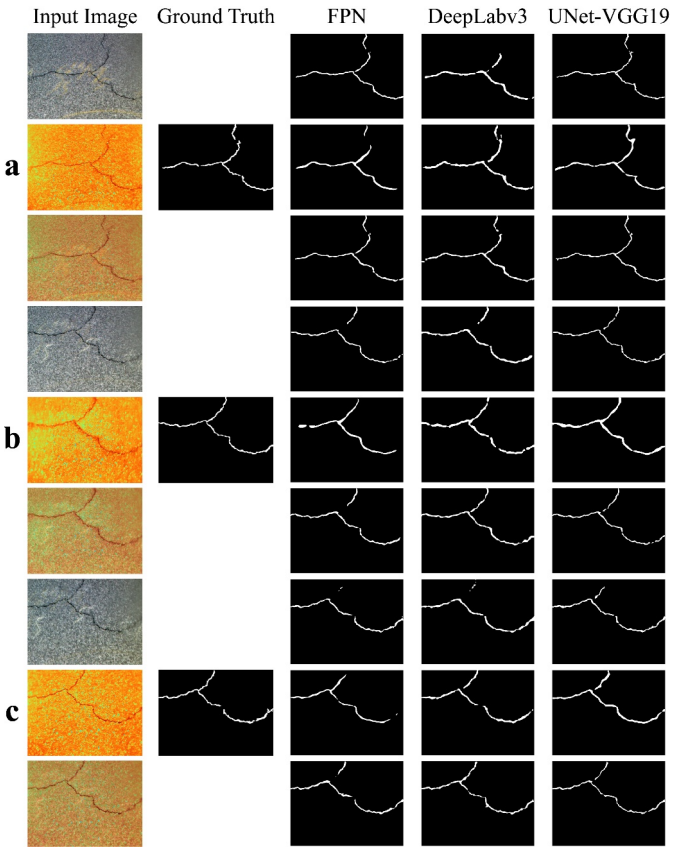

Fig. 13. Predictions of three segmentation models for different periods based on three types of images. (a) Morning, (b) noon, and (c) dusk.

The wet cracks tend to have a lower temperature and their areas become bluer, thus helping to detect cracks for the infrared image and fusion image.

In addition to various cracks and backgrounds, the dataset takes the periods into account. Different periods in a day have different temperatures and crack detection heavily depends on the temperature difference between cracks and backgrounds for the infrared image. Therefore, it needs to investigate the effect of the temperature on crack detection for the infrared image. Fig. 13 shows the predictions of three segmentation models for different periods based on three types of images. Although visible images in the three periods are almost identical, their corresponding infrared images and fusions are different (Fig. 3 and Fig. 13). The infrared image and fusion image in the morning and at dusk are similar and they are different at noon. It is observed that different periods in a day have no significant influence on the predictions of segmentation models for the infrared image and the fusion image.

Another thing worth noting is the indistinct edges of cracks and asphalt pavement surface. When cracks are similar to the background (e.g. pavement texture), it is difficult for segmentation models to detect cracks based on visible images (Fig. 12c and Fig. 13). The predictions of segmentation models possibly miss some parts of cracks in terms of the visible image. In comparison, the temperature difference would make cracks more distinguished in the infrared image and fusion image than the visible image (Fig. 12c and Fig. 13), thereby

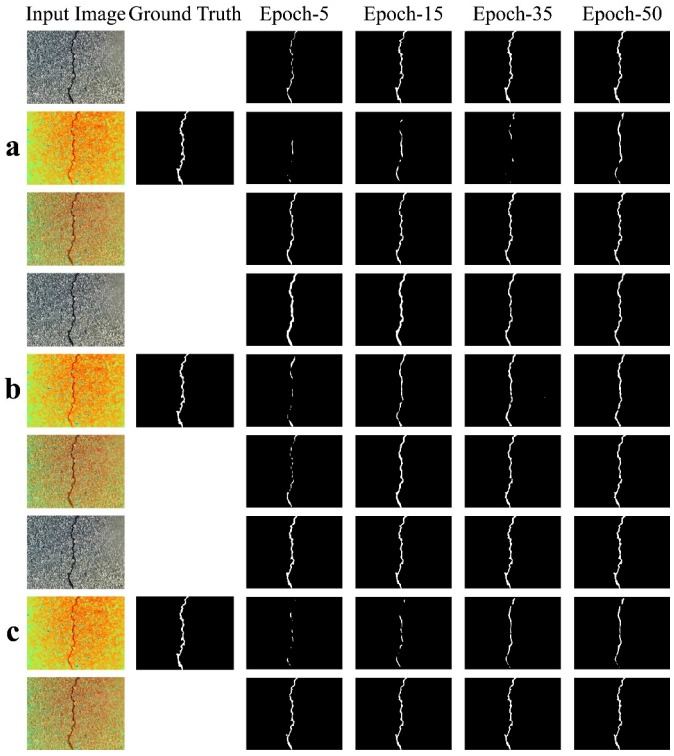

Fig. 14. Predictions of three segmentation models for different epochs based on three types of images. (a) FPN, (b) DeepLabv3, and (c) UNet-VGG19.

improving the quality of predictions. As shown in Fig. 12c and Fig. 13, the predictions of all segmentation models for the fusion image are better than the visible image. Furthermore, the predictions of the infrared image are similar to that of the visible image (Fig. 12c and Fig. 13), although they are worse than the visible image under other conditions. Therefore, when it is difficult to distinguish the edge of cracks and asphalt pavement surface, the fusion image is a better choice for crack detection.

To show the training process, the prediction of three segmentation models during different training epochs for three types of images is shown in Fig. 14. Predictions of the visible image and fusion image tend to become satisfactory and stable after 15 epochs for all segmentation models, while predictions of the infrared image need almost 50 epochs to stabilize and are still worse than predictions of the visible image and fusion image. Segmentation models have different performances on three types of images. All these models predict accurately for the visible image and fusion image, while DeepLabv3 has a better prediction of the infrared image than the other two segmentation models. For FPN, the fusion image allows the model to obtain a satisfying prediction faster (Fig. 14a in Epoch-5). But the predictions of Deeplabv3 are just the opposite and it is faster to predict accurately for the visible image. UNet-VGG19 gets a good prediction for the visible image and fusion image approximately at the same time.

V. Conclusion

CNN has become a powerful and efficient tool for detecting pavement cracks. One issue has been raised to find an accurate and efficient CNN model to keep a balance between accuracy and complexity in terms of crack detection. Another issue is the indistinct edges of cracks and asphalt pavement surface. This work aims to propose a robust crack detection method based on CNN and IRT. The main contribution of this work includes: (1) As generating gray-level information and temperature information, infrared thermography is used to better distinguish cracks, eliminating the effect of the indistinct edges between cracks and asphalt pavement surfaces. (2) Build an open benchmark dataset based on three types of images, including visible images, infrared images, and the fusion of visible and infrared images. This dataset also considers different conditions and periods, including single and multi cracks, thin and thick cracks, clean and rough backgrounds, light and dark backgrounds, and three periods in a day. (3) Seven CNN segmentation models are trained and evaluated on the aforementioned dataset to make a benchmark. (4) Evaluation metrics (accuracy, computational complexity, and model complexity) are used to have an overall evaluation of CNN segmentation models rather than only the accuracy metrics, thus finding an accurate and efficient CNN segmentation model.

Five classical CNN segmentation models and two UNet-based models are trained and evaluated on this dataset. The results show that the accuracy metrics and predictions of the visible image and fusion image are similar and even identical for all segmentation models, which are much better than that of the infrared image. When the background is rough (e.g. wet) or cracks are similar to the background (e.g. pavement texture), the fusion image is a better choice for crack detection as IRT distinguishes cracks by temperature differences. The different periods in a day have no significant influence on the predictions of segmentation models for the infrared image and the fusion image. Compared with the visible image, all segmentation models have a more stable performance for the fusion image. Among seven segmentation models, FPN could be the best model because of its high accuracy and low complexity.

Moreover, most of the images in this dataset include several cracks rather than a high percentage of cracks, which could be found by the ratio of crack pixels. In the future, more images would be collected from pavements with highly texted surfaces or with more complicated cracking patterns to enrich the dataset. And the size of the dataset would also increase by collecting and adding more images.

References

[1] Y. Hou *et al.*, "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, no. 6, pp. 845–856, 2020.

[2] Y. Hou *et al.*, "MobileCrack: Object classification in asphalt pavements using an adaptive lightweight deep learning," *J. Transp. Eng. B, Pavements*, vol. 147, no. 1, Mar. 2021, Art. no. 04020092.

[3] D. Kang *et al.*, "Hybrid pixel-level concrete crack segmentation and quantification across complex backgrounds using deep learning," *Autom. Construct.*, vol. 118, Oct. 2020, Art. no. 103291.

[4] *Standard Practice for Roads and Parking Lots Pavement Condition Index Surveys*, ASTM, West Conshohocken, PA, USA, 2009.

[5] A. Cubero-Fernandez, F. J. Rodriguez-Lozano, R. Villatoro, J. Olivares, and J. M. Palomares, "Efficient pavement crack detection and classification," *EURASIP J. Image Video Process.*, vol. 2017, no. 1, pp. 1–11, Dec. 2017.

[6] M. Salman *et al.*, "Pavement crack detection using the Gabor filter," in *Proc. ITSC*, Oct. 2013, pp. 2039–2044.
[7] Z. Liu, Y. Cao, Y. Wang, and W. Wang, "Computer vision-based concrete crack detection using U-Net fully convolutional networks," *Autom. Construct.*, vol. 104, pp. 129–139, Dec. 2019.
[8] B. Li, K. C. P. Wang, A. Zhang, E. Yang, and G. Wang, "Automatic classification of pavement crack using deep convolutional neural network," *Int. J. Pavement Eng.*, vol. 21, no. 4, pp. 457–463, Mar. 2020.
[9] Z. Fan, Y. Wu, J. Lu, and W. Li, "Automatic pavement crack detection based on structured prediction with the convolutional neural network," 2018, *arXiv:1802.02208*.
[10] N. D. Hoang, Q. L. Nguyen, and V. D. Tran, "Automatic recognition of asphalt pavement cracks using metaheuristic optimized edge detection algorithms and convolution neural network," *Autom. Construct.*, vol. 94, pp. 203–213, Oct. 2018.
[11] J. Huyan, W. Li, S. Tighe, Z. Xu, and J. Zhai, "CrackU-Net: A novel deep convolutional neural network for pixelwise pavement crack detection," *Struct. Control Health Monitor.*, vol. 27, no. 8, p. e2551, Aug. 2020.
[12] K. He and J. Sun, "Convolutional neural networks at constrained time cost," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Oct. 2015, pp. 5353–5360.
[13] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, P. Martinez-Gonzalez, and J. Garcia-Rodriguez, "A survey on deep learning techniques for image and video semantic segmentation," *Appl. Soft Comput.*, vol. 70, pp. 41–65, Sep. 2018.
[14] Y. Du, X. Zhang, F. Li, and L. Sun, "Detection of crack growth in asphalt pavement through use of infrared imaging," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2645, no. 1, pp. 24–31, Jan. 2017.
[15] Q. Zou, Y. Cao, Q. Li, Q. Mao, and S. Wang, "CrackTree: Automatic crack detection from pavement images," *Pattern Recognit. Lett.*, vol. 33, no. 3, pp. 227–238, 2012.
[16] C. Plati, P. Georgiou, and A. Loizos, "Use of infrared thermography for assessing HMA paving and compaction," *Transp. Res. C, Emerg. Technol.*, vol. 46, pp. 192–208, Jan. 2014.
[17] V. Vyas, V. J. Patil, A. P. Singh, and A. Srivastava, "Application of infrared thermography for debonding detection in asphalt pavements," *J. Civil Struct. Health Monitor.*, vol. 9, no. 3, pp. 325–337, Jul. 2019.
[18] H.-T. Bang, S. Park, and H. Jeon, "Defect identification in composite materials via thermography and deep learning techniques," *Composite Struct.*, vol. 246, Oct. 2020, Art. no. 112405.
[19] T. Yu, A. Zhu, and Y. Chen, "Efficient crack detection method for tunnel lining surface cracks based on infrared images," *J. Comput. Civil Eng.*, vol. 31, no. 3, May 2017, Art. no. 04016067.
[20] A. A. Oloufa, H. S. Mahgoub, and H. Ali, "Infrared thermography for asphalt crack imaging and automated detection," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 1889, no. 1, pp. 126–133, Jan. 2004.
[21] H. Zhao *et al.*, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2881–2890.
[22] T.-Y. Lin *et al.*, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2117–2125.
[23] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.
[24] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.
[25] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Oct. 215, pp. 234–241.
[26] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[27] K. He *et al.*, "Deep residual learning for image recognition," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[28] P. Yakubovskiy, "Segmentation models PyTorch," *GitHub Repository*, 2020. [Online]. Available: https://github.com/qubvel/segmentation_models.pytorch
[29] P. Huang, "The easiest implementation of fully convolutional networks," *GitHub Repository*, 2020. [Online]. Available: https://github.com/pochih/FCN-pytorch
[30] M. Alexandre, "UNet: Semantic segmentation with PyTorch," *GitHub Repository*, 2021. [Online]. Available: https://github.com/milesial/Pytorch-UNet
[31] N. P. Jouppi *et al.*, "In-datacenter performance analysis of a tensor processing unit," in *Proc. 44th Annu. Int. Symp. Comput. Archit.*, 2017, pp. 1–12.
[32] V. Sovrasov, "Flops counter for convolutional networks in PyTorch framework," *GitHub repository*, 2021. [Online]. Available: https://github.com/sovrasov/flops-counter.pytorch
[33] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," 2019, *arXiv:1912.01703*.
[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[35] S. Li, X. Zhao, and G. Zhou, "Automatic pixel-level multiple damage detection of concrete structure using fully convolutional network," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 34, no. 7, pp. 616–634, Jul. 2019.
[36] Y. Liu, J. Yao, X. Lu, R. Xie, and L. Li, "DeepCrack: A deep hierarchical feature learning architecture for crack segmentation," *Neurocomputing*, vol. 338, pp. 139–153, Apr. 2019.
[37] W. Liu *et al.*, "High-level semantic feature detection: A new perspective for pedestrian detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 5187–5196.
[38] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

Fangyu Liu received the B.S. and M.S. degrees in civil engineering from Tongji University, Shanghai, China, in 2017 and 2019, respectively.

Since 2019, he has been a Graduate Research Assistant with the Virginia Polytechnic Institute and State University (Virginia Tech). His research interests include fiber reinforced concrete and deep learning in structural health monitoring (crack detection and signal pattern recognition).

Jian Liu received the B.S. degree in traffic engineering from the Wuhan University of Science and Technology, Wuhan, China, in 2016, and the M.S. degree in infrastructure transportation engineering from the South China University of Technology, Guangzhou, China, in 2019.

Since 2019, he has been a Graduate Research Assistant with the Virginia Tech. His research interests include cement stabilized crushed rock, pavement structure, asphalt concrete design, machine learning, and deep learning in asphalt concrete design.

Linbing Wang received the B.S. degree in hydraulic engineering from Hohai University, Nanjing, China, in 1984, the M.S. degree in geotechnical engineering from Tongji University, Shanghai, China, in 1991, and the Ph.D. degree in civil engineering from the Georgia Institute of Technology (Georgia Tech), Atlanta, GA, USA, in 1998.

From 2000 to 2005, he was an Assistant Professor with Louisiana State University, Baton Rouge, LA, USA. In 2005, he joined the Virginia Polytechnic Institute and State University (Virginia Tech), Blacksburg, VA, USA, as an Associate Professor, where he became a Full Professor in 2010. He also worked as an Adjunct Professor at the University of Science and Technology Beijing. He is the author of one book and more than 230 journals and proceeding papers. His research interests include smart and sustainable technologies, energy harvesting, the IoT sensing networks, and health monitoring, innovative infrastructure assessment and predictive data analytics, material genome for multifunctional materials, multiple-scale characterization, modeling and simulation, pavement testing and mechanistic pavement design, transportation infrastructure preservation and risk management, and application of remote sensing and imaging techniques, and digital twins technology. He is a fellow of the American Society of Civil Engineers (ASCE) and the Engineering Mechanics Institute (EMI) of ASCE.