IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 27, 2019, accepted September 8, 2019, date of publication September 12, 2019, date of current version September 24, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2940767

Sample and Structure-Guided Network for Road Crack Detection

SIYUAN WU[1], JIE FANG[1,2], XIANGTAO ZHENG[1], (Member, IEEE), AND XIJIE LI[1]

[1]Key Laboratory of Spectral Imaging Technology CAS, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, China
[2]School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China

Corresponding author: Xiangtao Zheng (xiangtaoz@gmail.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61806193, Grant 61702498, and Grant 61772510, in part by the Young Top-Notch Talent Program of Chinese Academy of Sciences under Grant QYZDB-SSW-JSC015, in part by the Open Research Fund of the State Key Laboratory of Transient Optics and Photonics, in part by the Chinese Academy of Sciences under Grant SKLST2017010, in part by the CAS "Light of West China" Program under Grant XAB2017B26 and Grant XAB2017B15, and in part by the Xi'an Postdoctoral Innovation Base Scientific Research Project.

ABSTRACT As an indispensable task for traffic management department, road maintenance has attracted much attention during the last decade due to the rapid development of traffic network. As is known, crack is the early form of many road damages, and repair it in time can significantly save the maintenance cost. In this case, how to detect crack regions quickly and accurately becomes a huge demand. Actually, many image processing technique based methods have been proposed for crack detection, but their performances can not meet our expectations. The reason is that, most of these methods use bottom features such as color and texture to detect the cracks, which are easily influenced by the varied conditions such as light and shadow. Inspired by the great successes of machine learning and artificial intelligence, this paper presents a sample and structure guided network for detecting road cracks. Specifically, the proposed network is based on U-Net architecture, which remains the details from input to output by using skip connection strategy. Then, because the scale of crack samples is much smaller than that of non-crack ones, directly using the conventional cross entropy loss can not optimize the network effectively. In this case, the Focal loss is utilized to address the model optimization problem. Additionally, we incorporate the self-attention strategy into the proposed network, which enhances its stability by encoding the 2-order information among different local regions into the final features. Finally, we test the proposed method on four datasets, three public ones with labels and a photographed one without labels, to validate its effectiveness. It is noteworthy that, for the photographed dataset, we design a series of image processing strategies such as contrast enhancement to improve the generalization capability of the proposed method.

INDEX TERMS Road crack detection, neural network, representation capability, sample imbalance, structural information.

I. INTRODUCTION

Road maintenance is an important task for traffic management department, which has attracted much attention during the past decade since the relatively high construction cost of road network. Actually, crack is the early form of many road damages, and repair it in time can significantly reduce the maintenance cost [1]. The premise of repairing cracks is obtain its location and details accurately, which is a challenging problem. At present, the manual based road crack detection strategies still play an important role in road maintenance field. Except for their reliability, some limitations and disadvantages of manual based road crack detection are relatively severe, some of which are illustrated as follows: 1) the manual based road crack detection consumes too much human power, material resources and too long time. Specifically, the staff of maintenance department need to give realtime evaluation of road damage degree. However, in real world, the manual process is slow since the testing roads often have long distance and the job often requires lots of resources. 2) The manual evaluation results are usually inaccurate. Specifically, when the road crews generalize and evaluate cracks, the main base is the sense of naked eyes and subjective reflections of visual effects for the apparent characteristics such as length and width of the crack, which is relatively initiative and easy to result in large errors. 3) The manual

The associate editor coordinating the review of this manuscript and approving it for publication was Yan-Jun Liu.

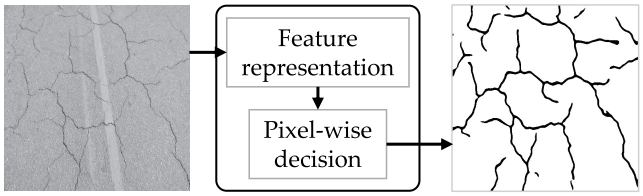


FIGURE 1. The framework of conventional road crack detection algorithms, which often contains two components, feature representation and pixel-wise decision. Specifically, this task aims to give each pixel in the image a specific attribution tags.

based road crack detection obstructs the normal road traffic. Specifically, when the staff of maintenance department inspect a road section, it needs to be closed. This will affect the normal use of the road and bring lots of inconveniences for the passing people and vehicles. 4) Manual detection has a certain dangers. Specifically, some roads are built on the steep and winding mountain roads, and the staff of maintenance department have certain personal safety risks when carrying out tasks.

To avoid the aforementioned disadvantages of manual based road crack detection, some traditional digital image processing technique [2], [3] based method have been proposed for automatical crack detection. For instance, the threshold analysis based methods [4], [5], the mathematical morphology based methods [6], [7], and the edge detection based methods [8], [9]. Most of the aforementioned methods are based on the optical and geometric assumptions for the properties of the crack images [5], [10], which are sensitive to the noise and their performances can not meet the application demands. Recently, with the rapid development of artificial intelligence and machine learning, deep neural networks especially deep convolutional neural networks have achieved significant performances on many computer vision tasks such as scene recognition [11], [12], salient object detection [13], [14], semantic segmentation [15], [16] and age prediction [17], [18]. As for road crack detection task, some CNN based pixel-wise or block-wise models [19], [20] have been proposed to improve the detection accuracy. Because of the strong feature representation capability, these CNN based methods have significant advantages compared to manual features based ones [21], [22]. For instance, Dorafshan *et al.* [21] found that CNN based methods are more robust to residual noises. However, several limitations still exist. For instance, most of them ignore the spatial relationships of crack regions among different pixels and sample scales of crack regions in the image. Specifically, spatial relationships of crack regions can not be characterized well with traditional independent pixel-level classification models, while which can provide extra auxiliary information for the final decision since the contrast is pivotal for crack detection. In addition, the scale of crack regions is much smaller than that of non-crack ones, and which can not effectively guide the model to possess the same sensitivity for crack and non-crack samples in testing phase.

To address the aforementioned problems, we propose a sample and structure guided network for road crack detection. Specifically, the proposed network is based on the popular U-Net [23], which has achieved huge successes for image transformation tasks because of its skip connection strategy. Then, considering the scale of crack samples is much smaller than that of non-crack ones, directly utilizing the conventional cross entropy loss can not optimize the network effectively, which may result in the ill-conditioned classifier and high leakage rate [24]. In this case, we use the Focal loss to optimize the network, which takes full consideration of the sample imbalance problem for crack images and utilizes the different penalty factors to train the model actually. Additionally, pixel-wise classification mechanism can not depict the structural relationships among different pixels in the image, which may result in the isolated noisy point in the predicted crack saliency map and improve the false drop rate. In this case, we incorporate the self- attention strategy into the proposed network, which enhances the stability of by encoding the 2-order interaction information among different local regions into the final features. Finally, differ from images in the public datasets, the photographed road images have not corresponding labels and they are often influenced by many environment conditions, such as light, shadow and lane line, which improve the detection complexity to a large extent. In order to generalize the model trained on public datasets to the photographed ones, we design a series of image processing strategies such as contrast enhancement, and apply them to the photographed images to improve the performances.

In summary, the contributions of this work can be listed as follows:

1) We propose a sample and structure guided network for road crack detection, which is based on pixel-wise classification mechanism. In addition, the proposed method considers the global structure and detailed texture information of the image simultaneously.
2) We utilize the Focal loss to guide the sample relationship learning, which addresses the network optimization problem by using different penalty factors for crack samples and non-crack ones.
3) We incorporate the self-attention mechanism into the network to guide the spatial structure learning, which alleviates the isolated noisy point problem by considering the relationships among different local regions in the image.
4) We propose a series of image processing techniques such as contrast enhancement to generalize the proposed algorithm to other open datasets, which improve its practical application value to a large extent.

The remainder of this paper is organized as follows. In section II, we introduce some existing methods for hyperspectral image classification, including unsupervised ones and supervised ones. Section III describes the proposed method. We report the experimental results in section IV and conclude the paper in section V.

II. RELATED WORKS

This section introduces some existing road crack detection methods, including the minimal path based methods, the image processing techniques (IPTs) based methods, the machine learning based methods and the deep learning based methods, which are introduced in subsection II-A, subsection II-B, subsection II-C and subsection II-D respectively.

A. MINIMAL BASED METHODS FOR ROAD CRACK DETECTION

The minimal path problem aims to find the best path among nodes in the graph, which has been applied to many tasks such as road crack detection. Kaul *et al*. [25] proposed an algorithm which works on much more general curve topologies with far fewer demands for initial input. Nguyen *et al*. [26] proposed a minimal path based algorithm for road crack detection, which considers the brightness and connectivity simultaneously to extract the characteristics of anisotropic cracks in the free-form path. Amhaz *et al.* [27] proposed an algorithm for automatic crack detection from 2D pavement images, which selects the endpoint and minimum path on local and global scales respectively. Even though these methods considers the features of crack pixels in the global view, their computational loads are too large and unsuitable for practical applications.

B. IPTs BASED METHODS FOR ROAD CRACK DETECTION

The current mainstream algorithms for road crack detection are all based on image processing techniques (IPTs), including the edge detection based methods and the histogram feature based method. Saar and Talvik [28] utilized Sobel operator to construct eight templates with different directions, and used them to extract the crack edges in road images, then they obtained the crack saliency map by combing the expansion processing technique of mathematical morphology with the iterated threshold algorithm. Velinsky and Kirschke [29] proposed to generate feature histograms of different levels in different regions, and extract features in various histograms to segment the images, which can achieve relatively satisfactory performances for images with obvious cracks.

C. MACHINE LEARNING BASED METHODS FOR ROAD CRACK DETECTION

With the development of machine learning, many approaches based on which have made great progressed for different applications [30]–[33]. Inspired by the successes of these methods, many feature extraction and pattern recognition based ones have been proposed for road crack detection [34]–[37]. Oliveira and Correia [35] proposed to utilize the mean and variance from unsupervised learning strategy to distinguish the crack blocks and non-crack ones. Cord and Chambon [36] proposed to utilize Adaboost strategy to choose the structural descriptors which can depict the crack images effectively, and then obtain the crack saliency map. Shi *et al*. [37] proposed to use a random forest based description method to depict cracks. Even though these methods have achieved relatively satisfactory performances, they are depend on the extracted features to a large extent, which limits their practical applications.

D. DEEP LEARNING BASED METHODS FOR ROAD CRACK DETECTION

With the significant improvement of the computational power of the hardware equipments, deep learning based methods have shown its advantageous performances on many vision tasks such as video object tracking [38], [39], object detection [40], image captioning [41], [42]. Recently, deep learning based methods have been successfully applied to damage and distress detection tasks. Cha *et al*. [20] proposed to divide the image into several blocks by sliding window strategy, and then successively judge whether the cracks exist in each block or not by CNN model. However, this method can only judge the block-level cracks, which is inaccurate. Zhang *et al.* [19] proposed to utilize CNN to judge whether each single pixel belongs to crack or not by using the local information of corresponding block, but they overestimated the crack width since ignored the spatial relationships among different pixels. Zhang *et al*. [43] proposed to use CNN to predict the label of each pixel in the image. However, this method needs to extract features by using manual feature descriptors and the CNN model only acts as a classifier. Additionally, their network architecture is closely related to the size of input image, which hinders the promotion of this method.

III. PROPOSED METHOD

This section details the proposed sample and structure guided network. Specifically, subsection III-A introduces the overview of the network, subsection III-B introduces the sample guidance learning strategy, subsection III-C introduces the structure guidance learning strategy. Additionally, subsection III-D introduces the designed image enhancement series for public and photographed datasets in detail.

A. OVERVIEW

We consider the road crack detection task as a pixel-wise classification one, and utilize a U-Net [23] based model, whose architecture is shown in Fig. 2, to finalize it. Specifically, the skip connection strategy can preserve the detailed texture information from raw images to corresponding predicted crack saliency maps. Then, we utilize the popular Focal loss [24] based sample guidance strategy to optimize the network, which can alleviate the ill-conditioned classifier from imbalanced samples. Additionally, we incorporate the self-attention [44] based structure guidance strategy into the network, which can avoid the isolated noisy point problem to a large extent. Last but not the least, we propose a series of image enhancement strategies to generalize the proposed algorithms to other open datasets more conveniently and accurately.

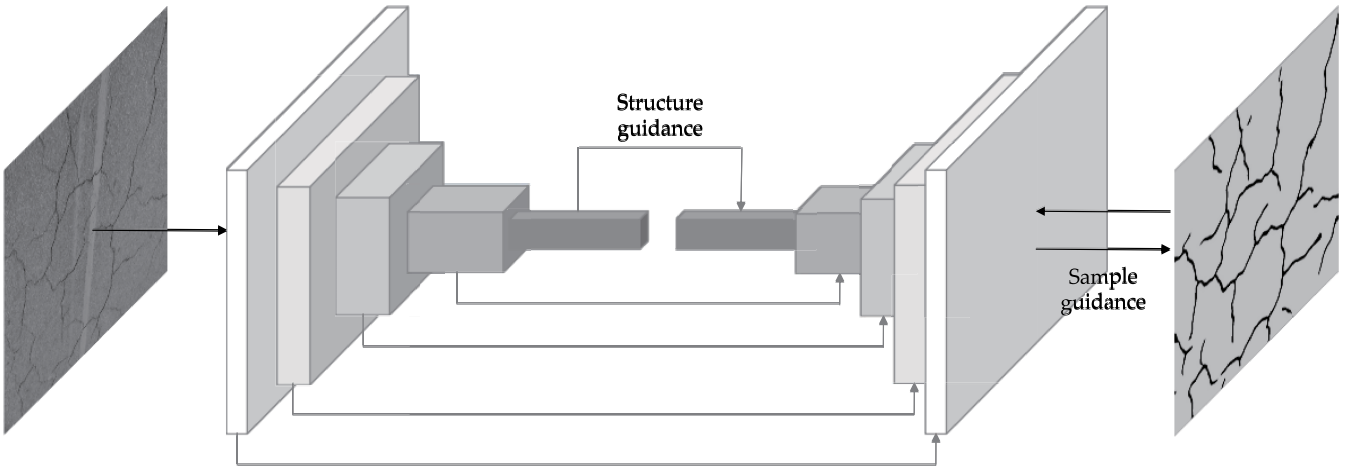


FIGURE 2. The architecture of the proposed sample and structure guided network. Based on the popular U-Net, the proposed network is equipped with spatial guidance and sample guidance modules to improve its effectiveness.

B. SAMPLE GUIDANCE LEARNING

This subsection introduces the sample guidance learning strategy. As is known, deep learning based methods obtain the intrinsic attributes from training sets themselves in an absolutely data-driven way. In other words, the distribution and quality of the data directly affect the performance of the model. As for road images, the scale of crack samples is much smaller than that of non-crack ones, which is actually a severe sample imbalance problem. In this case, directly use the conventional MSE loss or cross entropy loss, which gives the same penalty factors to each sample in the image, may results in the model bias problem. Specifically, because most samples of the training set belong to non-crack category, the trained model is insensitive for the crack samples and tend to predict all samples to non-crack category, which influences the performance especially brings to the high miss rate to a large extent. In these cases, we utilize the Focal loss, which gives larger penalty factor to category with smaller scale and smaller penalty factor to category with larger scale, to optimize the proposed network effectively. The Focal loss is defined as Equation 1,

$$\mathcal{L} = -\frac{1}{WH}\sum_{w=1}^{W}\sum_{h=1}^{H}\left(\alpha y_{(w,h)}\left(1-\hat{y}_{(w,h)}\right)^{\eta}\log\hat{y}_{(w,h)} + (1-\alpha)\, y_{(w,h)}\hat{y}_{(w,h)}^{\gamma}\log\left(1-\hat{y}_{(w,h)}\right)\right), \quad (1)$$

where W and H represents the width and height of the image respectively. $y_{(w,h)}$ and $\hat{y}_{(w,h)}$ represents the label and predicted saliency score of $(w, h)_{th}$ pixel in the image respectively. Additionally, α and η are two hyperparameters, which are used to guide the sample learning. Specifically, the hyperparameter pair with larger α and smaller η means the greater emphasis on crack samples.

Actually, even though cracks are discussed at length in this paper, which are not only defect that the matters to inspectors. In this case, for the purpose of popularizing our method to other more complex road maintenance systems conveniently, we have formed a variant of focal loss and which can classify the data with more than two classes. The definition of our expanded focal loss is shown as Equation 2,

$$\ell = -\frac{1}{WH}\sum_{w=1}^{W}\sum_{h=1}^{H}\sum_{c=1}^{C}\alpha_c y_{(w,h)}^{c}\left(1-\hat{y}_{(w,h)}^{c}\right)^{\eta_c}\log\hat{y}_{(w,h)}^{c}, \quad (2)$$

where C is the class number of the dataset. $y_{(w,h)}$ is the one-hot label of $(w, h)_{th}$ pixel in the image and $y_{(w,h)}^{c}$ represents its c_{th} element. Specifically, $y_{(w,h)}^{c} = 1$ $(w, h)_{th}$ pixel belongs to c_{th} class, and otherwise $y_{(w,h)}^{c} = 0$. Besides, $\hat{y}_{(w,h)}$ represents the predicted score vector of $(w, h)_{th}$ pixel, and which has the same size with c_{th} class. In addition, α_c and η_c are two parameters to control the optimization process of c_{th} class.

C. STRUCTURE GUIDANCE LEARNING

This subsection introduces the structure guidance strategy. Actually, the relationships among different pixels are vital for the pixel-wise classification tasks [45]. As is known, convolutional kernels can consider the relationships of pixels in a local region, but they can not depict the relationships among different pixels in a global view [44]. Specifically, an important function of convolution operation is to represent the present pixel with ones in its surroundings, and which can consider the interactions of pixels in a specific local region well. However, pixels with longer physical distances can not be used to represent each other by convolutional kernels even their characteristics are similar since their limited receptive fields. For road crack detection task, both the relationships in a local region and ones in a global view can contribute to the final results. Specifically, the contrast but not the absolute intensity is the main evidence for crack detection because of the variable photographed conditions such as light and shadow. The relationships among different pixels in a local region can provide some auxiliary contrast information, compared to classify each pixel in an independent way. Additionally, the global structure priors are also

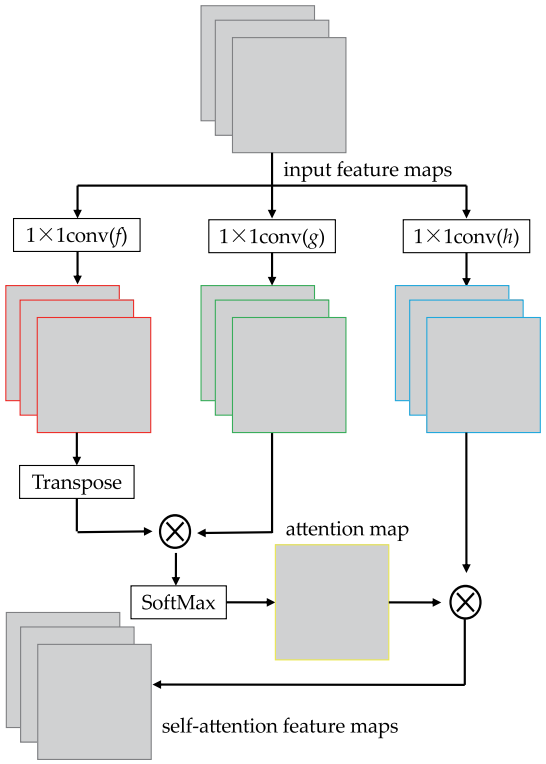


FIGURE 3. The flowchart of the Self-Attention mechanism, which mainly contains three subbranches. Specifically, a subbranch (h) is used to obtain the routine convolutional features of the image, and two other subbranches (f and g) are used to calculate the attention map, which contains interaction information of each paired local blocks. Then the attention map is multiplied to the routine feature map to obtain the final self-attention feature maps.

important for the crack detection. The relationships among different local regions can effectively avoid the isolated noisy points or isolated noisy blocks in the predicted crack saliency map, because they can provide the sufficient directional priors of texture in a global view. In this case, we incorporate the self-attention mechanism, which is effective for the representation of long range dependency, into the conventional U-Net, to improve its performance. It is noteworthy that, the self-attention module is only followed the last layer in encode part because of the computational load. The flowchart of the self-attention mechanism is shown in Fig. 3, which is described in details as follows.

The feature maps from the last encode layer $\mathbf{x} \in \mathbf{R}^{D\times N}$ are first fed into two spaces f and g to obtain the new feature maps $f(\mathbf{x}) = \mathbf{W}_f\mathbf{x} + \mathbf{b}_f$ and $g(\mathbf{x}) = \mathbf{W}_g\mathbf{x} + \mathbf{b}_g$. Then the relationships among different local regions in the image can be calculated with Equation 3,

$$r_{(j,i)} = \frac{e^{s_{(i,j)}}}{\sum_{i=1}^{N} e^{s_{(i,j)}}}, \tag{3}$$

where $s_{(i,j)} = f(\mathbf{x}_i)^T g(\mathbf{x}_j)$, and $r_{(j,i)}$ demotes the extent to which the model attends to the i_{th} block when representing the j_{th} block. Additionally, the self-attention feature is denoted as $\mathbf{o} \in \mathbf{R}^{D\times N}$, and whose element can be calculated by using Equation 4,

$$\mathbf{o}_j = \sum_{i=1}^{N} r_{(j,i)}\,(h(\mathbf{x}_i)), \tag{4}$$

where $h(\mathbf{x}) = \mathbf{W}_h\mathbf{x} + \mathbf{b}_h$. Finally, the output of the self-attention module is formulated as Equation 5,

$$\mathbf{y} = \mathbf{x} + \nu\mathbf{o}, \tag{5}$$

where ν is a hyperparameter to balance the original features and the self-attention features. Additionally, in order to improve the representation capability of self-attention mechanism [44], we add three bias $\mathbf{b}_f$, $\mathbf{b}_g$, and $\mathbf{b}_h$ to $f(\cdot)$, $g(\cdot)$, and $h(\cdot)$ respectively.

D. IMAGE ENHANCEMENT STRATEGIES

This subsection introduces the designed image enhancement strategy, including the enhancement strategy for the public datasets and that for the photographed datasets, which are introduced as follows.

1) IMAGE ENHANCEMENT STRATEGY FOR PUBLIC DATASETS

For reasons of skid resistance and construction cost, although the surface of normal road is relatively flat, the road details are not very smooth because of the gaps among different small gravels. These gaps form the discrete small shadow regions in the collected road images including ones in the public and photographed datasets, which have the similar gray intensity with the cracks and the relatively obvious boundaries with the surrounding regions. In these cases, the boundaries of the shadow regions are often misclassified as cracks, which influences the performance of the method to a certain extent.

To address the aforementioned problem, we design a weighed filtering strategy to alleviate the influence of the small shadow regions. It is obvious that, the scale of shadow regions due to the gaps among different gravels is much smaller than that of the normal regions, and the filtering strategy can represent the shadow regions with the combination ones, which can decrease the contrast between them and improve the performance further. The flowchart of the designed weighted filtering strategy is shown in Fig. 4, which is described in details as follows. N median filters $f_i^{m_i}(\cdot)_{1\le i\le N}$ with different sizes $m_i = 2^i + 1$ are first applied to the original road image $\mathbf{I}$ to obtain the filtered image series $\mathbf{F}_i = f_i^m(\mathbf{I})_{1\le i\le N}$, then the enhanced image can be obtained by summing these filtered image series with different weight coefficients, which is defined as Equation 6,

$$\mathbf{I}^f = \sum_{i=1}^{N} \gamma_i\mathbf{F}_i, \tag{6}$$

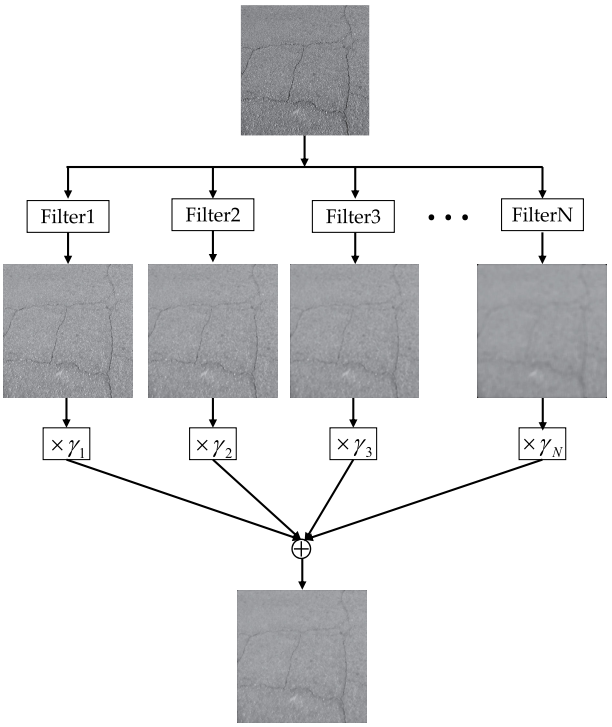


FIGURE 4. The flowchart of the designed image enhancement mechanism for public datasets. Specifically, we a) apply a series of median filter kernels with different sizes to obtain corresponding filtered image series, b) multiply different weight coefficients to corresponding filtered images, and c) sum the weighted filtered images as the final processing result.

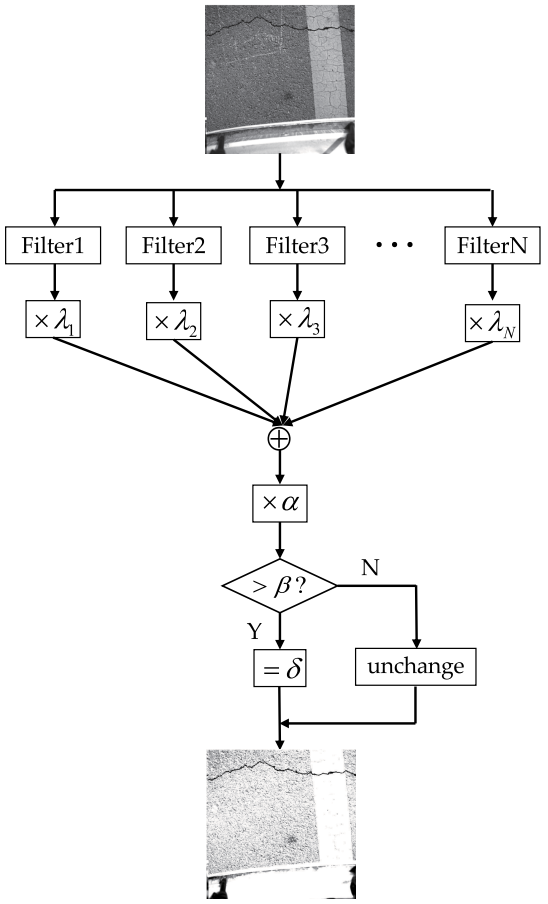


FIGURE 5. The flowchart of the designed image enhancement mechanism for photographed datasets.

where $\gamma_i = \frac{2^{-i}}{\sum_{i=1}^{N} 2^{-i}}$. As can be seen from Fig. 4, bigger median kernels can remove more noisy points while smaller ones can remain more detailed textures. In this case, our enhancement strategy can reconcile the noise removal and crack texture remaining of road images simutaneously while a simple median filter can only address ones of these problems. In addition, we place more emphasis on texture information because of its importance for crack detection, which can be seen from the definition of γ_i.

2) IMAGE ENHANCEMENT STRATEGY FOR PHOTOGRAPHED DATASETS

As for the photographed datasets, besides shadows from gaps among different gravels, guideposts such as lane line interfere the crack detection severely. These interferences can not be tackled simply by the aforementioned weighted filtering strategy since the scale of these samples are large and the contrasts between them and original road are significantly evident.

In these cases, we design a threshold based contrast improvement strategy to address the aforementioned problem, whose flowchart is shown in Fig. 5. Specifically, we first utilize the aforementioned weighted filtering strategy to suppress the influences of the noise and small shadow regions caused by the gaps among different gravels. After the weighted filtered image $\mathbf{I}^f$ is obtained, we multiply it with a factor $\alpha(\alpha > 1)$ to obtain an intensity-enhanced image $\mathbf{I}^\alpha$. As for $(w, h)_{th}$ pixel $\mathbf{I}^\alpha_{(w,h)}$ in $\mathbf{I}^\alpha$, if $\mathbf{I}^\alpha_{(w,h)} > \beta$, we set $\mathbf{I}^\alpha_{(w,h)} = \delta$ and then we can obtain the contrast enhanced image $\mathbf{I}^c$. Because the intensity of the crack pixel $\mathbf{I}^c$ is often low, the intensity of $\alpha\mathbf{I}^c$ also remains in a relatively low level and often can not surpass β. Additionally, the intensities of other pixels $\mathbf{I}^o$, including ones in normal road and lane line regions are relatively high, so $\alpha\mathbf{I}^o$ are likely to surpass β and be set to δ. In other words, the designed strategy enlarges the contrast between crack and normal regions, while suppresses the contrast between normal and lane line regions. In these cases, the designed threshold based contrast improvement strategy can improve the detection performance by alleviating the effects such as guideposts.

It is noteworthy that, the similar mean intensity of images in training and testing sets can contribute the final results, which should be adjusted when generalizing the model trained on public datasets to photographed datasets.

IV. EXPERIMENTS

This section details the experiments, including datasets, experimental settings, evaluation metrics, contrasting methods and experimental results, which are introduced

TABLE 1. Important specifications of camera used to photograph CrackPV dataset.

Specifications	GS3-U3-123S6M-C
Maximum resolution/Frame frequency	$4096 \times 3000@30fps$
Pixel sensor	Sony IMX253 CMOS $1.1''$
Pixel size	$3.1\mu m$
ADC	$14bits/10bits$
Transmission rate	$5Gbits/s$

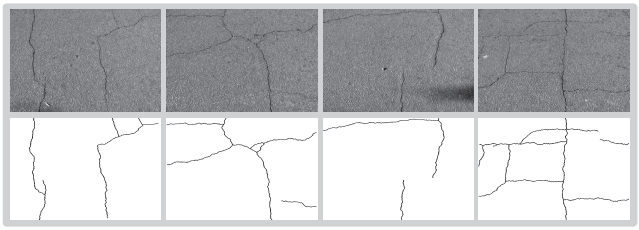


FIGURE 6. Some samples of CrackTree200 dataset, images in the first and second rows are raw pavement ones and the corresponding groundtruths.

in subsection IV-A, subsection IV-B, subsection IV-C, subsection IV-C, and subsection IV-E respectively.

A. DATASETS

In order to validate the effectiveness of the proposed method, we test it on four datasets, including CrackTree200 dataset [46], ALE dataset [27], CrackForest dataset [37] and CrackPV dataset. Specifically, CrackTree200, ALE and CrackForest are three datasets for scientific research. CrackPV is the dataset we photographed on moving vehicle. The material used in the pavement structures of these four datasets are all asphalt, which increases the detection difficulty since its low contrast in crack and normal regions. In addition, because the type of camera plays a major role in the metrics of CNNs, some important specifications of camera to photograph our CrackPV dataset are shown in Table 1. Finally, the four aforementioned datasets are introduced in details as follows.

1) CRACKTREE200 DATASET

CrackTree200 dataset contains 206 pavement images with the fixed size of 800 × 600 pixels. Additionally, interferences such as shadows and noises improve the detection difficulty of CrackTree200 dataset. Some samples of CrackTree200 dataset are shown in Fig. 6, from which we can see that the surfaces of these images are not smooth as we expected, and the small gaps in the image will influence the detection results.

2) CRACKFOREST DATASET

CrackForest dataset contains 118 images with the fixed size 480×320 pixels, and each image has corresponding manually labeled groundtruth. Some samples of CrackForest dataset are shown in Fig. 7, from which we can see that the light conditions of different images and even different regions in the same image. In addition, images in CrackForest dataset

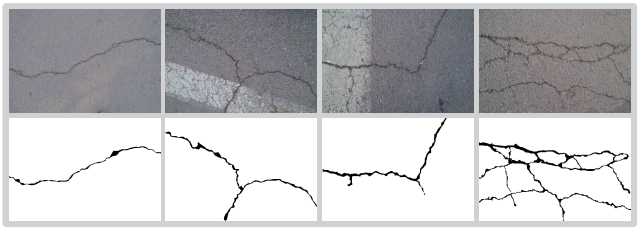


FIGURE 7. Some samples of CrackForest dataset.

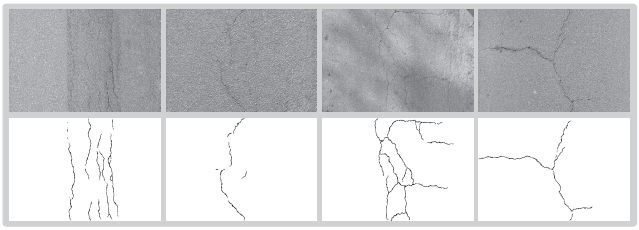


FIGURE 8. Some samples of ALE dataset.

FIGURE 9. Some samples of CrackPV dataset.

contain noises such as oil spots and water stains, all of the these factors increase its difficulty.

3) ALE DATASET

ALE dataset contains three subsets actually, including ESAR, LCMS and Aigle-RN, which are respectively obtained by three imaging systems, named ESAR, LCMS and Aigle-RN. Specifically, ESAR corresponds to a static acquisition with no controlled lighting, LCMS uses laser and Aigle-RN utilizes stroboscopic lights. In addition, ESAR contains 15 images with fully annotated labels. LCMS contains five pixel-wise annotated groundtruths. Aigle-RN contains 38 images with pixel-wise labels. Some samples of ALE dataset are shown in Fig. 8.

4) CRACKPV DATASET

CrackPV dataset is the one we photographed on moving vehicle with the speed of 80km per hour, compared to the public datasets, the photographed one is more complex and challenging. Specifically, the motion blurring interference, the shadow interference, the guideposts interference and other controllable and uncontrollable factors such as exposure intensity and weather condition will influence the image quality to a large extent. Some samples of CrackPV dataset are shown in Fig. 9.

B. EXPERIMENTAL SETTINGS

In this subsection, we introduce the experimental settings, including data partition, hyperparameter settings and testing platform, which are described as followings.

1) DATA PARTITION

For three public datasets, we choose 50% samples as the training set and leave the other half as the testing one. For CrackPV set, we utilize the popular transfer learning strategy to use the model trained on CrackTree200 dataset to predict the crack saliency map. Because the scales of three public datasets are relatively small, which are insufficient to train a robust and effective model. In this case, we utilize the data augmentation strategy introduced in [23] to expand the training sets.

2) TESTING PLATFORM

The algorithms are implemented with Pytorch, and the testing platform is X99UD4 of GIGABTYE, GPU (8G×8) of Titan X. For each of the contrasting algorithms, we train the model through mini-batch SGD with batch size 4. The initial learning rate is set to 5×10^{-5}, the momentum is set to 0.9 and each model is trained for 10 epoches. The image enhancement procedures we designed in this paper are implemented with Matlab on CPU platform.

C. EVALUATION METRICS

To verify the superiority of the proposed method, we use three common metrics to evaluate its performances, including precision (P), recall (R), and F-measure (F_β). Specifically, P and R are defined in Equation 7,

$$P = \frac{|M \cap G|}{|M|}, R = \frac{|M \cap G|}{|G|}, \tag{7}$$

where M is the binary mask obtained from the crack saliency map through a specific threshold, G is the corresponding manual annotated label map. Besides, as a weighted harmonic mean of P and R with a non-negative β, F_β is formulated as Equation 8,

$$F_\beta = \frac{\left(1+\beta^2\right) \cdot P \cdot R}{\beta^2 \cdot P + R}, \tag{8}$$

where β is a hyperparameter to balance P and R. As is suggested in [47], β^2 is set to 0.3 to emphasize the importance of precision.

D. CONTRASTING METHODS

Three methods are used as the contrasting ones to verify the superiority of this work, including fully convolutional network (FCN) [48], Dilated convolutional network (DiCN) [49] and the U-Net [23]. FCN is the ground-breaking work for applying deep convolutional neural network to semantic segmentation task in an end-to-end way. DiCN expands the receptive field through adding holes with different sizes to convolutional kernels but not shrinks the feature maps, which can maintain more texture information. U-Net utilizes the skip-connection strategy to propagate more texture information from shallow layers to deeper ones, which has achieved huge success for salient object detection task.

TABLE 2. Ablation experimental results on CtrackTree200 dataset.

/	SaGui	WeiFil	StGui	P(%)	R(%)	F_β(%)
Ablation	–	–	–	0	0	0
	√	–	–	14.04	88.63	17.42
	√	√	–	15.06	89.04	18.63
	√	√	√	**15.47**	**90.11**	**19.03**

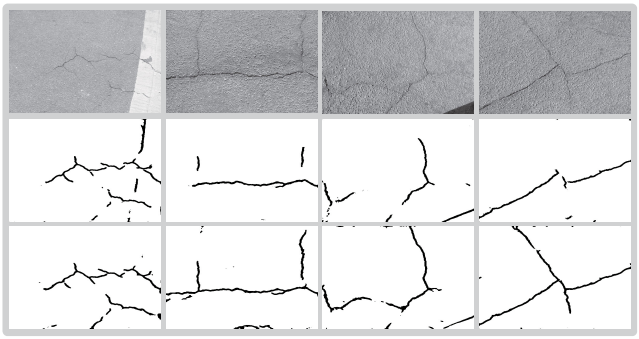


FIGURE 10. The visualized results of ablation experiment on CrackTree200 dataset, raw images are shown in the first row, the results of U-Net+SaGui are shown in the second row, and the results of U-Net+SaGui+WeiFil+StGui are shown in the third row.

E. EXPERIMENTAL RESULTS

This subsection reports the experimental results, including ablation experimental results, contrasting experimental results, and experimental results on open datasets. Specifically, the ablation experiment is to validate the effectiveness of each proposed component, the contrasting experiment is to verify the superiority of the proposed method compared to the existing ones, and the experiment on open datdaset is to demonstrate the generalization and practicability of the proposed method.

1) ABLATION EXPERIMENTAL RESULTS

Compared to the conventional U-Net for saliency detection task, the proposed method contains three differences, the weighted-filtering strategy, the self-attention alike based structure guidance and the Focal loss based sample guidance modules. We incorporated them into the conventional U-Net successively to validate their effectiveness, and the ablation results are shown in Table 2, where SaGui, WeiFil and StGui represents the sample guidance, weighted filtering and structure guidance strategy respectively. Additionally, some visualized results are shown in Fig. 10.

From Table 2 we can see that, each of the proposed components has contributed to the improvement of the detection performance. Specifically, U-Net equipped with conventional cross entropy loss even can not detect any crack regions, which is because the ill-conditional problem from the severe imbalanced samples in training set. When the Focal loss

TABLE 3. Contrasting experimental results on three datasets.

Dataset	Method	P(%)	R(%)	F_β(%)
CrackTree200	FCN	5.53	87.92	7.06
	DiCN	10.55	77.24	13.18
	U-Net	14.04	88.63	17.42
	SSGN	**15.47**	**90.11**	**19.03**
ALE	FCN	14.19	**96.64**	17.67
	DiCN	14.68	82.14	18.11
	U-Net	34.89	94.66	40.84
	SSGN	**41.88**	93.99	**48.03**
CrackForest	FCN	25.31	**77.42**	29.96
	DiCN	29.79	72.36	34.47
	U-Net	40.71	75.27	45.54
	SSGN	**43.30**	76.23	**48.09**

based sample guidance strategy is applied, the detection performances have improved a lot. Besides, weighted filtering and structure guidance strategies have enhanced the performance further, which obtain 1.21% and 0.40% increments in terms of F_β respectively.

From Fig. 10 we can see that, the proposed three components can alleviate the influence of shadow (images in the first column), alleviate the fracture condition of predicted cracks (images in the second the third columns), and reduce the omission rate (images in the fourth column) to a certain extent. The reasons contain two aspects: 1) the weighted filtering strategy decreases the contrast among normal and shadow regions, 2) and the self-attention based structure guidance learning strategy improves the detection performances through considering the 2-order information of different blocks in the image. In general, all these phenomenons can validate the effectiveness of the proposed components.

Incidentally, performance of the proposed method seems inferior to some related works [21], [43] if only from the perspective of metric indictors. Actually, the profound reasons for this situations are different evaluation strategies but not the algorithms themselves. Specifically, our evaluation strategy is based on pixel-level and some others are based on block-level measurements. Additionally, for CrackTree200 dataset, the width of crack in groundtruth is narrower than that in corresponding original road image. In testing stage, the predicted crack saliency line is often wider than that in groundtruth since all contrasting methods used in this paper are based on the pixel-level classification mechanism, which results in a huge impact on pixel-level based evaluation strategy but a smaller one on block-based evaluation strategy.

2) CONTRASTING EXPERIMENTAL RESULTS

This part reports the contrasting experimental results with four algorithms, including three existing ones and ours, on three public datasets, which are shown in Table 3. Incidentally, SSGN in Table 3 represents the proposed sample and structure guided network.

Experimental results on CrackTree200 dataset are shown in the first block of Table 3. From which we can see that,

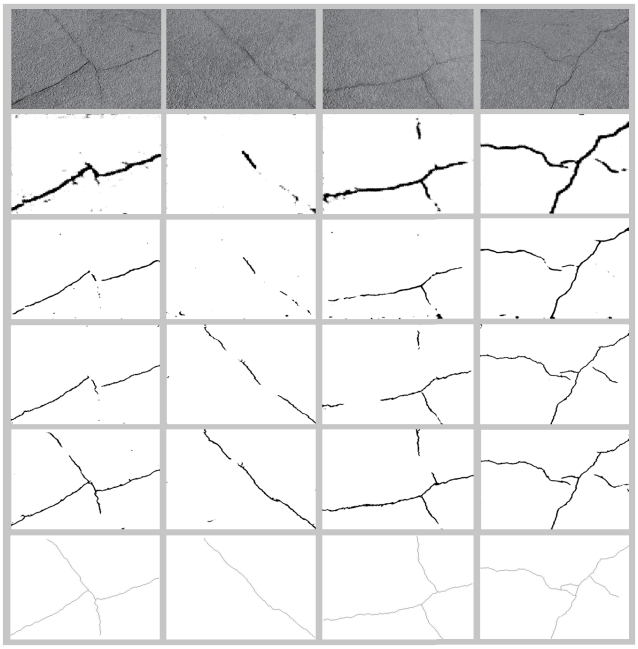


FIGURE 11. The visualized results of CrackTree200 dataset. Images in the first and last rows are raw images and the corresponding groundtruth. Additionally, Images in the second to fifth rows are corresponding saliency maps predicted by FCN, DiCN, U-Net and SSGN respectively.

DiCN achieves better performances than FCN. Specifically, which obtains 5.03% improvement in terms of precision and 6.12% improvement in terms of F_β, while the recall rate decreases from 87.92% to 77.24%. The reason is that, DiCN enlarges the receptive field through hole convolutional kernels, which remains more detail information and improves the precision. However, hole convolutional kernels can not depict the intrinsic information of independent pixel well, while this attribute information of independent pixel is very important for crack detection since the crack regions are often relatively small. In this case, the leakage rate of DiCN is higher and the recall rate is correspondingly lower than those of FCN.

Some visualized results on CrackTree200 dataset are shown in Fig. 11. From which we can see that, the crack saliency map predicted by DiCN is thinner than that by FCN, while the fracture phenomenon of the former is more severe than that of the latter.

Experimental results on ALE dataset are shown in the second block of Table 3. From which we can see that, U-Net significantly improves the detection performance, compared to DiCN. Specifically, it achieves 20.12%, 12.52%, 22.73% improvements in terms of precision, recall and F-measure respectively, which demonstrate the superiority of U-Net architecture for crack detection task sufficiently. On one hand, even though convolutional kernels can enlarge the receptive field with out shrinking feature maps, they often bring to many noises since violate the local similarity rule in images. On the other hand, U-Net utilizes the skip connection strategy to remain the spatial details from the raw images to corresponding saliency maps, while it uses the traditional but

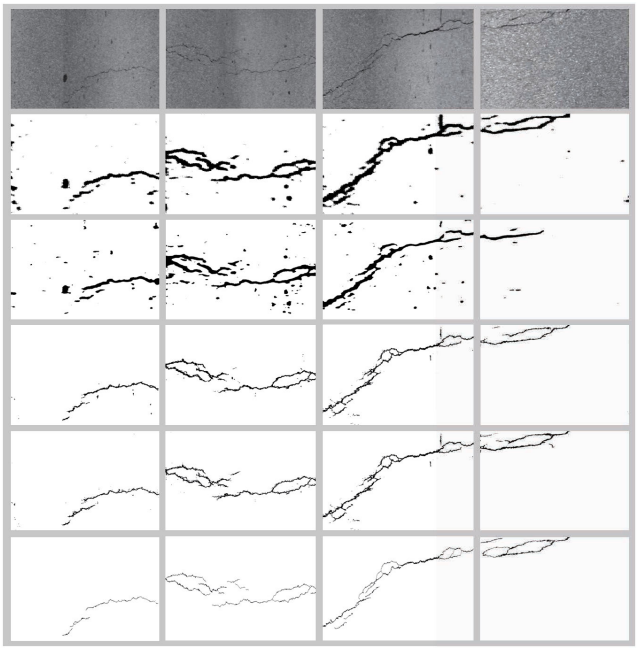


FIGURE 12. The visualized results of ALE dataset. Images in the first and last rows are raw images and the corresponding groundtruth. Additionally, Images in the second to fifth rows are corresponding saliency maps predicted by FCN, DiCN, U-Net and SSGN respectively.

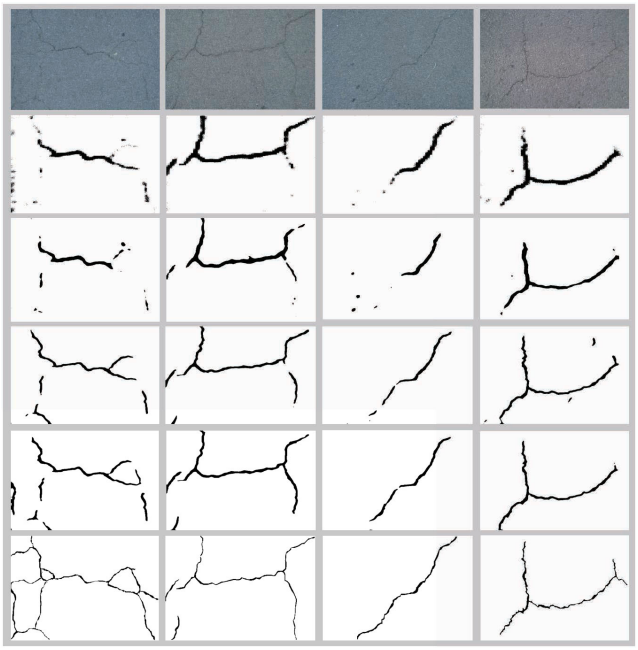


FIGURE 13. The visualized results of CrackForest dataset. Images in the first and last rows are raw images and the corresponding groundtruth. Additionally, Images in the second to fifth rows are corresponding saliency maps predicted by FCN, DiCN, U-Net and SSGN respectively.

not the hole convolutional kernels to alleviate the noises. Both the aforementioned reasons contribute to the superiority of U-Net.

Some visualized results on ALE dataset are shown in Fig. 12. From which we can see that, compared to saliency maps predicted by DiCN, ones predicted by U-Net are more accurate, which have more thinner crack skeletons and fewer isolated noisy points and small noisy blocks.

Experimental results on CrackForest dataset are shown in the third block of Table 3. From which we can see that, the proposed SSGN achieves better performances than U-Net. Specifically, which achieves 2.59%, 0.96% and 2.55% improvements in terms of precision, recall rate and F_β respectively. The main reasons contain two aspects, the proposed self-attention based structure guidance module encodes interactions among different local regions in the image to the final predicted saliency map, which enhances the global structure representation. Additionally, the designed weighted filtering strategy alleviates interference of noises and small shadow regions from gaps among different gravels.

Some visualized results on CrackForest dataset are shown in Fig. 13. From which we can see that, compared to U-Net, the proposed SSGN addresses the crack fracture problem to a large extent. The results of first and second images can illustrate this point apparently.

The above analysis demonstrates the effectiveness and superiority of our proposed sample and structure guided network. It is noteworthy that, as shown in Table 3, performances of the same model on different datasets are significantly different. For instance. FCN achieves 5.53% and 25.31% in terms of precision on CrackTree200 and CrackForest datasets respectively, which have a huge gap. Besides the intrinsic characteristics such as intensity and contrast of images in different datasets, another important reason for this situation is the different labeling strategies. Specifically, crack width in groundtruth of CrackTree200 dataset remain a single pixel regardless of the actual crack width, while which in groundtruth of CrackForest dataset varies with the actual crack width in the corresponding original image. Additionally, crack width in predicted saliency map varies with the actual crack of the corresponding original images since all contrasting methods are based on pixel-level classification mechanism. In these cases, metrics such as precision have shown a significant difference even for the similar actual prediction errors on two datasets.

3) EXPERIMENTAL RESULTS ON CRACKPV DATASET

Compared to the experimental results on public datasets, the performances on practical ones are more important since the original intention of designing the algorithm is to assist the road maintenance task. In this case, we test the proposed method equipped with conventional corrosion operations on CrackPV, an open dataset we photographed on moving vehicle. It is noteworthy that we transfer to use the model trained on CrackTree200 dataset to predict the crack saliency map of images in CrackPV dataset, because the latter does not have the corresponding manually annotated pixel-wise labels. Additionally, some visualized results are shown in Fig. 14. From which we can see that, most of the crack regions have been detected, and the interference such as lane lines and gaps

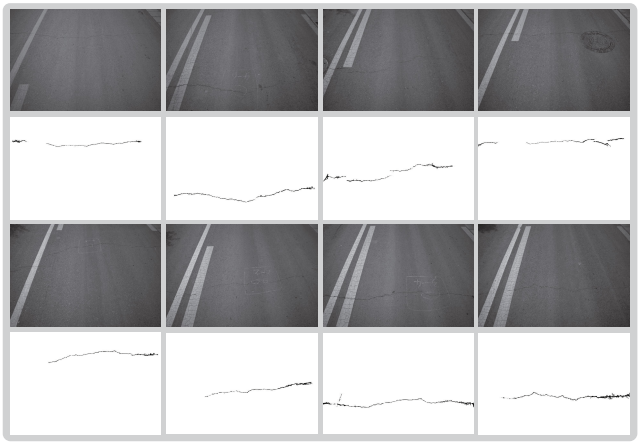


FIGURE 14. The visualized results on CrackPV dataset, images in the first and third rows are raw road images, while ones in the second and fourth rows are corresponding predicted binary saliency map.

among different gravels are suppressed well in the predicted crack saliency map. All these positive effects are benefited from the designed threshold based contrast improvement strategy, which sufficiently enhances the robustness of the algorithm. Generally speaking, combined with appropriate image processing techniques, the proposed sample and structure guided network can be generalized to open datasets and achieve relatively satisfactory performances.

V. CONCLUSION AND FUTURE WORK

In this paper, we propose a sample and structure guided network for road crack detection, which considers the task as a pixel-wise classification one and can obtain the crack saliency map from the raw road image directly. Specifically, we utilize the Focal loss to guide the sample relation learning, which addresses the optimization problem from imbalanced data. Then, we incorporate the self-attention mechanism into the network to guide the spatial structure learning, which alleviates the isolated noisy point problem. Additionally, we propose a series of image enhancement strategies to generalize the proposed method to other open datasets, which improves its practical application value to a large extent. Finally, experimental results on three public and a photographed datasets validate the robustness, effectiveness and superiority of the proposed algorithm.

ACKNOWLEDGMENT

(Siyuan Wu and Jie Fang contributed equally to this work.)

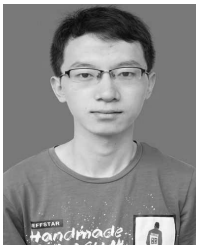
REFERENCES

[1] R. S. Lim, H. M. La, Z. Shan, and W. Sheng, "Developing a crack inspection robot for bridge maintenance," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2011, pp. 6288–6293.
[2] X. Li, K. Liu, Y. Dong, and D. Tao, "Patch alignment manifold matting," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 7, pp. 3214–3226, Jul. 2018.
[3] X. Li, K. Liu, and Y. Dong, "Superpixel-based foreground extraction with fast adaptive trimaps," *IEEE Trans. Cybern.*, vol. 48, no. 9, pp. 2609–2619, Sep. 2018.
[4] H. Oliveira and P. L. Correia, "Automatic road crack segmentation using entropy and image dynamic thresholding," in *Proc. 17th Eur. Signal Process. Conf.*, Aug. 2009, pp. 622–626.
[5] S. Chambon and J.-M. Moliard, "Automatic road pavement assessment with image processing: Review and comparison," *Int. J. Geophys.*, vol. 2011, Jun. 2011, Art. no. 989354.
[6] N. Tanaka and K. Uematsu, "A crack detection method in road surface images using morphology," in *Proc. MVA*, Nov. 1998, pp. 17–19.
[7] J. Tang and Y. Gu, "Automatic crack detection and segmentation using a hybrid algorithm for road distress analysis," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, Oct. 2013, pp. 3026–3030.
[8] H. Zhao, G. Qin, and X. Wang, "Improvement of canny algorithm based on pavement edge detection," in *Proc. 3rd Int. Congr. Image Signal Process.*, vol. 2, Oct. 2010, pp. 964–967.
[9] Q. Li and X. Liu, "Novel approach to pavement image segmentation based on neighboring difference histogram method," in *Proc. Congr. Image Signal Process.*, vol. 2, May 2008, pp. 792–796.
[10] Y. Hu and C.-X. Zhao, "A novel LBP based methods for pavement crack detection," *J. Pattern Recognit. Res.*, vol. 5, no. 1, pp. 140–147, 2010.
[11] Y. Yuan, J. Fang, X. Lu, and Y. Feng, "Remote sensing image scene classification using rearranged local features," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 3, pp. 1779–1792, Mar. 2019.
[12] J. Fang, Y. Yuan, X. Lu, and Y. Feng, "Robust space-frequency joint representation for remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, to be published.
[13] K.-J. Hsu, Y.-Y. Lin, and Y.-Y. Chuang, "Weakly supervised salient object detection by learning a classifier-driven map generator," *IEEE Trans. Image Process.*, vol. 28, no. 11, pp. 5435–5449, Nov. 2019.
[14] L. Han, X. Li, and Y. Dong, "Convolutional edge constraint-based U-net for salient object detection," *IEEE Access*, vol. 7, pp. 48890–48900, 2019.
[15] 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. 2017.
[16] J. Fang and X. Cao, "GAN and DCN based multi-step supervised learning for image semantic segmentation," in *Proc. 1st Chin. Conf. Pattern Recognit. Comput. Vis. (PRCV)*, Guangzhou, China, Nov. 2018, pp. 28–40.
[17] J. Fang, Y. Yuan, X. Lu, and Y. Feng, "Muti-stage learning for gender and age prediction," *Neurocomputing*, vol. 334, pp. 114–124, Mar. 2019.
[18] K. Jhang and J. Cho, "CNN training for face photo based gender and age group prediction with camera," in *Proc. Int. Conf. Artif. Intell. Inf. Commun. (ICAIIC)*, Feb. 2019, pp. 548–551.
[19] L. Zhang, F. Yang, Y. D. Zhang, and Y. J. Zhu, "Road crack detection using deep convolutional neural network," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 3708–3712.
[20] Y.-J. Cha, W. Choi, and O. Büyüköztürk, "Deep learning-based crack damage detection using convolutional neural networks," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 32, no. 5, pp. 361–378, 2017.
[21] S. Dorafshan, R. J. Thomas, and M. Maguire, "Comparison of deep convolutional neural networks and edge detectors for image-based crack detection in concrete," *Construct. Building Mater.*, vol. 186, pp. 1031–1045, Oct. 2018.
[22] K. Makantasis, E. Protopapadakis, A. Doulamis, N. Doulamis, and C. Loupos, "Deep convolutional neural networks for efficient vision based tunnel inspection," in *Proc. IEEE Int. Conf. Intell. Comput. Commun. Process. (ICCP)*, Sep. 2015, pp. 335–342.
[23] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, Munich, Germany, Oct. 2015, pp. 234–241.
[24] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2980–2988.
[25] V. Kaul, A. Yezzi, and Y. C. Tsai, "Detecting curves with unknown endpoints and arbitrary topology using minimal paths," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 10, pp. 1952–1965, Oct. 2012.
[26] T. S. Nguyen, S. Begot, F. Duculty, and M. Avila, "Free-form anisotropy: A new method for crack detection on pavement surface images," in *Proc. 18th IEEE Int. Conf. Image Process.*, Sep. 2011, pp. 1069–1072.
[27] R. Amhaz, S. Chambon, J. Idier, and V. Baltazart, "Automatic crack detection on two-dimensional pavement images: An algorithm based on minimal path selection," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 10, pp. 2718–2729, Oct. 2016.

[28] T. Saar and O. Talvik, "Automatic asphalt pavement crack detection and classification using neural networks," in *Proc. 12th Biennial Baltic Electron. Conf.*, Oct. 2010, pp. 345–348.
[29] S. A. Velinsky and K. R. Kirschke, "Design considerations for automated pavement crack sealing machinery," in *Applications of Advanced Technologies in Transportation Engineering*. Reston, VA, USA: ASCE, 1991, pp. 76–80.
[30] Z. Yu and H.-S. Wong, "Quantization-based clustering algorithm," *Pattern Recognit.*, vol. 43, no. 8, pp. 2698–2711, 2010.
[31] T. Gao, Y.-J. Liu, L. Liu, and D. Li, "Adaptive neural network-based control for a class of nonlinear pure-feedback systems with time-varying full state constraints," *IEEE/CAA J. Autom. Sinica*, vol. 5, no. 5, pp. 923–933, Sep. 2018.
[32] S. Wu, H.-S. Wong, and Z. Yu, "A Bayesian model for crowd escape behavior detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 1, pp. 85–98, Jan. 2014.
[33] L. Liu, Y.-J. Liu, and S. Tong, "Fuzzy based multi-error constraint control for switched nonlinear systems and its applications," *IEEE Trans. Fuzzy Syst.*, vol. 27, no. 8, pp. 1519–1531, Aug. 2019.
[34] P. Delagnes and D. Barba, "A Markov random field for rectilinear structure extraction in pavement distress image analysis," in *Proc. Int. Conf. Image Process.*, vol. 1, Oct. 1995, pp. 446–449.
[35] H. Oliveira and P. L. Correia, "Automatic road crack detection and characterization," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 1, pp. 155–168, Mar. 2013.
[36] A. Cord and S. Chambon, "Automatic road defect detection by textural pattern recognition based on AdaBoost," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 27, no. 4, pp. 244–259, 2012.
[37] Y. Shi, L. Cui, Z. Qi, F. Meng, and Z. Chen, "Automatic road crack detection using random structured forests," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 12, pp. 3434–3445, Dec. 2016.
[38] S. Wu, X. Li, and X. Lu, "Robust object tracking via diverse templates," in *Proc. Int. Conf. Comput., Inf. Telecommun. Syst. (CITS)*, Jul. 2016, pp. 1–5.
[39] X. Zhang, Y. Yuan, and X. Lu, "Deep object tracking with multimodal data," in *Proc. Int. Conf. Comput., Inf. Telecommun. Syst. (CITS)*, Jul. 2016, pp. 1–5.
[40] Y. Zhang, Y. Yuan, Y. Feng, and X. Lu, "Hierarchical and robust convolutional neural network for very high-resolution remote sensing object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5535–5548, Aug. 2019.
[41] X. Li, A. Yuan, and X. Lu, "Vision-to-language tasks based on attributes and attention mechanism," *IEEE Trans. Cybern.*, to be published.
[42] A. Yuan, X. Li, and X. Lu, "3G structure for image caption generation," *Neurocomputing*, vol. 330, pp. 17–28, Feb. 2019.
[43] A. Zhang, K. C. Wang, B. Li, E. Yang, X. Dai, Y. Peng, Y. Fei, Y. Liu, J. Q. Li, and C. Chen, "Automated pixel-level pavement crack detection on 3D asphalt surfaces using a deep-learning network," *J. Comput.-Aided Civil Infrastruct. Eng.*, vol. 32, no. 10, pp. 805–819, 2017.
[44] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," 2018, *arXiv:1805.08318*. [Online]. Available: https://arxiv.org/abs/1805.08318
[45] F. Shen, R. Gan, S. Yan, and G. Zeng, "Semantic segmentation via structured patch prediction, context CRF and guidance CRF," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1953–1961.
[46] 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.
[47] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 733–740.
[48] 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.
[49] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*. [Online]. Available: https://arxiv.org/abs/1511.07122

SIYUAN WU is currently with the Key Laboratory of Spectral Imaging Technology CAS, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, Shaanxi, China.

His current research interests include image processing, video object tracking, and stereo matching.



JIE FANG received the B.S. degree from the School of Electronic Engineering, Xidian University, Xi'an, Shaanxi, China, in 2015. He is currently pursuing the Ph.D. degree in signal and information processing techniques with the Key Laboratory of Spectral Imaging Technology CAS, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, and the University of Chinese Academy of Sciences, Beijing, China.

His research interests include artificial intelligence, machine learning, and image understanding.

XIANGTAO ZHENG received the M.Sc. and Ph.D. degrees in signal and information processing from the Chinese Academy of Sciences, Xi'an, Shaanxi, China, in 2014 and 2017, respectively.

He is currently an Assistant Professor with the Key Laboratory of Spectral Imaging Technology CAS, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences. His main research interests include computer vision and pattern recognition.

XIJIE LI is currently with the Key Laboratory of Spectral Imaging Technology CAS, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, Shaanxi, China.

His research interests include image processing, design of spectrometer, and spectral data analysis.