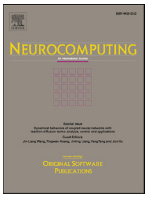
Neurocomputing 424 (2021) 193–204

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



Brief papers

Distribution equalization learning mechanism for road crack detection

Jie Fang[a,b,*], Bo Qu[a], Yuan Yuan[c]

[a] *Key Laboratory of Spectral Imaging Technology CAS, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, Shaanxi, PR China*

[b] *University of Chinese Academy of Sciences, Beijing 100049, PR China*

[c] *Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, PR China*

ARTICLE INFO

Article history:
Received 3 August 2019
Revised 11 December 2019
Accepted 13 December 2019
Available online 19 December 2019

Communicated by Dr. Ruiping Wang

Keywords:
Road crack detection
Salient object detection
Imbalanced sample
Ill conditioned classifier

ABSTRACT

Visual-based road crack detection becomes a hot research topic over the last decade because of its huge application demands. Road crack detection is actually a special form of salient object detection task, whose objects are small and distribute randomly in the image compared to the traditional ones, which increase the difficulty of detecting. Most conventional methods utilize bottom information, such as color, texture, and contrast, to extract the crack regions in the image. Even though these methods can achieve satisfactory performances for images with simple scenarios, they are easily interfered by some factors such as light and shadow, which may decrease the detection result directly. Inspired by the competitive performances of deep convolutional neural networks on many visual tasks, we propose a distribution equalization learning mechanism for road crack detection in this paper. Firstly, we consider the crack detection task as a pixel-level classification and use a U-Net based architecture to finalize it. Secondly, the occurrence probability of crack and non-crack are so different, which results in the ill-conditioned classifier and undesirable detection performance, especially the high false detection rate. In this case, we propose a weighted cross entropy loss term and a data augmentation strategy to avoid influence from imbalanced samples through emphasizing the crack regions. Additionally, we propose an auxiliary interaction loss, and combine it with the popular self-attention strategy to alleviate the fracture situations through considering relationships among different local regions in the image. Finally, we tested the proposed method on three public and challenging datasets, and the experimental results demonstrate its effectiveness.

© 2019 Elsevier B.V. All rights reserved.

1. Introduction

The maintenance of highways has important impact on efficiency in the use of road, since its construction cost is usually significantly high. An important premise of highway maintenance is to obtain the damaged section, including subsidence areas, fracture zones, regions with many cracks. However, field investigation is always costly, regardless of human resources, material resources or financial resources.

In these cases, many Structural Health Monitoring (SHM) techniques have been proposed recently, especially the visual-based ones [1,2]. These methods primarily use image processing techniques (IPTs) to tackle the problem, which have a significant advantage that almost all superficial defects can be detected [3,4]. Abdel-Qader et al. utilized four edge detection methods, Canny edge detector, Sobel edge detector, Fast Fourier Transform and Fast Haar Transform, to detect concrete cracks [5], who found that FHT is the best solution for the task. Even though these edge detection based methods have achieved relatively satisfactory performances, they are easily affected by the noises from lighting and distortion [6].

With the rapid development of machine learning and artificial intelligence, convolutional neural networks are successfully applied to many visual understanding tasks, such as image recognition [7], object detection [8], semantic image segmentation [9], etc. As for the crack detection task, Cha et al. proposed a regional classification based method using deep convolutional neural network [10], which has improves the robustness for noises because of the strong feature representation capability of deep neural network. However, the detection result of this method is relatively coarse since it only gives each 256 × 256 block a binary label, and the cracks can not be depicted effectively.

To address the aforementioned issues, we consider the crack detection as a salient object detection task. Specifically, each pixel in an image can be classified into two categories, crack or not,

[*] Corresponding author at: University of Chinese Academy of Sciences, Beijing 100049, PR China.

E-mail address: jackfang713508@gmail.com (J. Fang).

https://doi.org/10.1016/j.neucom.2019.12.057
0925-2312/© 2019 Elsevier B.V. All rights reserved.

and we can obtain the crack regions directly without any post-processing, such as the sliding window procedure in [10]. However, there will be a severe situation if directly applying conventional salient object detection network to crack datum: *data imbalance problem*. As is known, traditional salient object detection networks are actually pixel-level classification ones, and the datum distribution is very important for their performances. As for road crack images, the scale of cracks is much smaller than that of non-crack, which is certain to result in the ill-posed classifier and high undetected ratio further. The details about this point are demonstrated in the Section 4. In this paper, we utilize two strategies to address the data imbalance problem, one is the data augmentation, and the other is the weighted classification loss. Specifically, our data augmentation strategy is not simply applying procedures such as rotation, translation to images, but increasing the scale of crack sample to balance the data distribution. The augmentation strategy is based on the reasonable assumption that, if a pixel in an image belongs to the crack category, the pixels in its neighborhood also belong to the crack category. The neighbourhood range is a hyperparameter, which influences the sophistication degree of the predicted results. Additionally, even though the data augmentation strategy increases the sample scale of crack regions, data imbalance problem still exists. In this case, we further utilize the weighted cross entropy loss to address this issue, which gives different weight coefficients to pixels with different labels. Specifically, the classification loss of crack is larger than that of non-crack, which avoids ill-posed classifier through emphasizing the category with smaller sample scale. Finally, we design an auxiliary loss to avoid the fracture situation through considering interactions among different regions in the image, which reflects the entire structure information of the cracks in the image and improves the performance of the model further. In summary, the contributions of the proposed method can be concluded as follows:

1. We propose a truncated expansion based data augmentation strategy to relieve the sample imbalance issue, which is realized by increasing the small-scale-category samples. The strategy is based on the assumption that, if a pixel in an image belongs to the crack category, the pixels in its neighborhood also belong to the crack category.
2. We propose a weighted cross entropy loss term to avoid the ill-posed classifier issue through emphasizing the crack samples, which does not need setting hyperparameters manually like focal loss. Specifically, the proposed loss term gives different weight coefficients to pixels with different categories. Detailedly, the weight for crack is larger than that of non-crack.
3. We propose an auxiliary interaction loss term, and combine it with the popular self-attention strategy to alleviate the fracture situation through considering interactions among different regions in the image. Specifically, the interaction loss term improves the effectiveness of the method by encoding the crack structure information into the feature representation.

The rest of this paper is organized as follows. In Section 2, we introduce some works related to our method. Section 3 describes the proposed method. We report the experimental results in Section 4 and conclude the paper in Section 5.

2. Related works

This section details some related works for road crack detection. First of all, we introduce some detection method in Section 2.1. Additionally, because our proposed method is based on salient object detection ones, some salient object detection methods are introduced in Section 2.2. Finally, since the most challenging point in crack detection task is data imbalance, we introduce some approaches for this issue in Section 2.3.

2.1. Crack detection

Crack detection is a challenging task in structural health monitoring (SHM) field. At the very beginning, vibration based structural systems using numerical method conjugations are popular [11,12]. Even though those methods contribute to the development of SHM, they still have several challenges for monitoring large scale civil infrastructures because of various uncertainties and nonuniformly distributed environmental effects, among other factors. Additionally, although several works have had large SHM performed to cover large scale structures, dense instrumentations for environmental effects are required. Finally, confirming whether the collected data contains structural damage, noisy signals, sensory system malfunction or a combination of them is not easy before checking the sensing systems and structures in person. Recently, with the rapid development of image processing techniques, a large number of vision based crack detection methods are proposed, including handcrafted feature based ones [13–15] and deep learning based ones [10,16,17]. Yeum et al. proposed a study for detecting cracks using image processing techniques combined with a sliding window technique, they found the potential of image processing techniques very well [13]. Even though the testing samples contain many crack-like features because of the rusty surface, the unnecessary features were removed effectively, and strong crack-like features were extracted by Frangi filter and Hessian matirx based edge detector [18]. Cha et al. proposed a vision based method using deep convolutional neural network (CNN) for detecting concrete cracks without calculating the defect features [10]. Because of the strong feature representation capability, CNN based methods are relatively robust to lighting and shadows, compared to the handcrafted ones.

2.2. Salient object detection

Existing salient object detection methods can be mainly divided into two branches, handcrafted based approaches and learning based ones. Itti et al. proposed to measure saliency through center-surround contrast of intensity, color and orientation features [19]. Fu et al. proposed to utilize normalized graph cut (Ncut) for salient object detection [20]. Gong et al. proposed to postpone the propagation to difficult regions and advance the propagation to simple regions to improve the detection performance [21]. Zhou et al. proposed a novel framework which includes localized estimation, spatiotemporal refinement, and saliency update parts to improve the detection results [22]. Xie et al. defined the salient object detection task in a Bayesian framework and predicted visual salient regions through a likelihood probability [23]. Cheng et al. used global contrast of a region with its corresponding contexts to calculate saliency [24]. Additionally, background prior is also usually applied to handcrafted methods [25–28], where the fundamental hypothesis is that the boundary regions of images more likely belong to background. Even though these hand-crafted saliency methods are efficient and effective, and they have achieved satisfactory performances on relative simple scenarios, they are not robust in tackling complex scenarios, and are easy to be interfered by the noises from uncertain conditions, such as lighting and shadow. Recently, learning based methods [29–31] have attracted more attention for salient object detection task. These methods utilize robust detectors trained with pixel-level annotations to detect saliency regions automatically [32,33]. Additionally, with the rapid progress of machine learning and artificial intelligence, deep convolutional neural network based methods have achieved competitive performances on salient object detection tasks. For example, Wang et al. proposed to detect saliency through training a DNN-G and a DNN-L network for global search and local estimation respectively [33]. Zhao et al. proposed to model superpixel saliency through multi-context CNNs considering both global and local information of the image [34]. More recently, Lee et al. used both low-level and high-level features for salient object detection through a unified deep learning framework, which encodes the low-level distance map to enhance the salient object detection performance [35]. Liu et al. proposed Structure Inference Net, which considers scene-level context and instance-level relationships simultaneously to enhance the feature representation, and obtains more accurate results [36] further.

2.3. Data imbalance

Data imbalance problem is a ubiquitous issue in pattern recognition field, which affects the property of the recognition systems to a large extent. Previous works on data imbalance issue can be divided into two branches, data-based methods and algorithm-based ones [37]. Data-based methods manipulate the category representations in the original dataset through either oversampling the minority categories or undersampling the majority categories to make the sampled data distribution balanced. Even though these methods are effective and convenient to tackle, they usually change the distribution of original data, and consequently bring some disadvantages. Through increasing the scale of training set, oversampling makes the training process costly, *i.e.* time and computational load. Additionally, undersampling may lose useful information of the majority category data potentially [38]. In these cases, some algorithm-based methods are proposed, which modify the training procedure to improve the sensitivity of the classifier to minority categories directly. Li et al. gave more importance to minority category samples through setting weights with Adaboost during the training process of the learning perceptron [39]. As for neural networks, Kukar et al. found that, the incorporation of costs into the loss function can improve the recognition performance on imbalanced data [40]. Chung et al. proposed a novel Cost-Sensitive loss function, which uses a regression loss to replace the traditional softmax one [41]. Raj et al. proposed a loss function, which gives equal importance to majority and minority categories [42]. Differing to these existing methods, we propose a new strategy to address the data imbalance problem, which utilizes data augmentation and algorithm modification techniques simultaneously.

3. Proposed method

This section details the proposed distribution equalization learning mechanism for road crack detection. Specifically, Section 3.1 introduces the overview of the mechanism which mainly includes three components, image preprocessing, network and data augmentation strategy. Section 3.2 introduces the details of image preprocessing. Section 3.3 introduces the network architecture and loss function of the proposed method. Section 3.4 details our data augmentation strategy.

3.1. Overview

We consider the road crack detection task as a domain transformation one, from raw road image domain to crack saliency domain. In this case, we utilize a salient object detection method to tackle this problem, which aims to attach a specific label, crack or otherwise, to each pixel in the image. Motivated by the competitive performances of U-Net for structural prediction tasks due to its skip connection style, we design our mechanism based on it. Additionally, because the road crack detection is sensitive to noises, we apply a series of preprocessing techniques to images before putting them to the network. Finally, a threshold subsection function is used to obtain the crack label maps from the predicted crack saliency maps. The flowchart of the proposed mechanism is shown in Fig. 1, which mainly consists of three components, image preprocessing component, deep neural network component and data augmentation component.

3.2. Image preprocessing

As is known, road is constructed with many pebbles, which makes the road surface not smooth even though there is no crack. In this case, the false-out ratio will increase if directly use raw road image to detect the crack regions. To address this issue, we propose a median-filtering based preprocessing strategy. Specifically, we apply three median kernels with different sizes (k=3, 7, and 15, respectively) to each raw image, and obtain three corresponding filtered images. Then, we obtain the final processed image through calculating mean value of these three filtered images, and the process can be formulated as Eq. (1).

$$I_p = \frac{1}{N}\sum_{n=1}^{N} F_n(I_r), \tag{1}$$

where I_p represents the final processed image, I_r represents the original raw image. F_n represents the n_{th} median filtering operation. N is the number of filtering operation, which is set to 3 in this paper. On one hand, applying median filtering operation to original raw images can avoid the interferences of noises, which can be seen in Fig. 2. On the other hand, the filtered images with different median kernels actually reflect different perceptual information of the image. Specifically, the larger ones depict global structure information while the smaller ones depict local texture information. Obviously, their combination can represent the crack details and avoid interferences from noises well simultaneously.

3.3. Network

This subsection details the inner architecture and loss function of the proposed mechanism. As is known, convolutional neural networks (CNNs) have achieved significant performances on many computer vision tasks, such as image recognition and face recognition. These tasks have a common characteristic, which encodes image into a discriminative feature vector. Because of the inherent reasonable advantages of receptive conception and the data-driven supervised learning strategy, the learned feature representations with CNNs are more discriminative than handcrafted ones. However, as for structural prediction tasks, CNN-based methods often give fragmented outputs due to the use of large-receptive-field convolution kernels and many pooling layers.

3.3.1. Network architecture

Differing from traditional recognition task, image to image transformation tasks, focus more on local texture information but not the global discriminative one, especially road crack detection. Because of the aforementioned issues, a structure preserving network is significantly vital for road crack detection, which must preserve sufficient detailed texture information from input image to predicted saliency map. Recently, U-Net and its variants have shown significant performances on several structural prediction tasks. Specifically, during the forward propagation, U-Net concatenates the lower-layer texture-sensitive features with corresponding deeper-layer semantic-sensitive features if their sizes are same. Through this skip-connection procedure, both semantic information and texture information are preserved well to the predicted label map, compared to the traditional fully convolutional neural networks. The inner architecture of the U-Net is shown in Table 1.

Additionally, even though U-Net can propagate detailed texture information of the image from lower to deeper layers well, it does

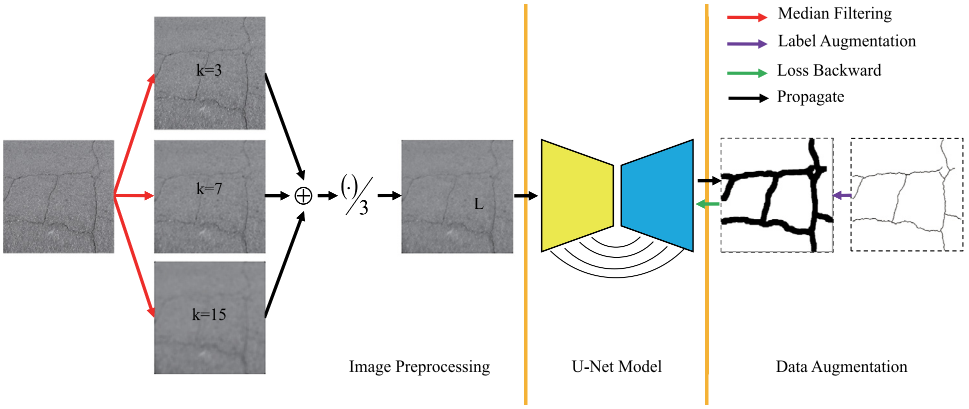


Fig. 1. The overview of the proposed method, which mainly consists of three components, image preprocessing component, deep neural network (U-Net) component and data augmentation component.

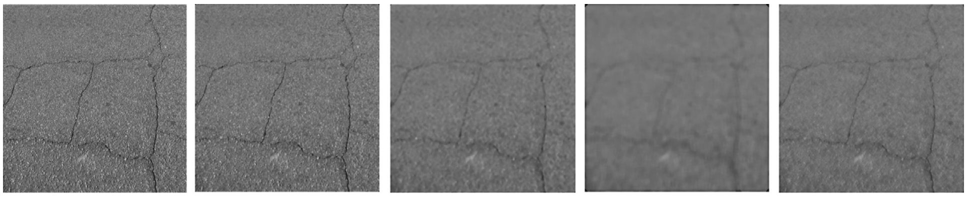


Fig. 2. The visualized results of applying median kernels with different sizes to an original raw image. From left to right, images in the figure represents original raw image, filtered images by median kernels with size 3, 7, and 15, combination of three filtered images respectively. From which we can see that, the noises are held up and the cracks are remained well in the combination image of three filtered ones.

Table 1
U-Net architecture.

Layer	C/S	Input	Input size	Layer	C/S	Input	Input size
data	1	/	/	*conv*10 (output)	1	*conv*9$_2$	512 × 512 × 64
*conv*1$_1$	64	*data*	512 × 512 × 3	*conv*9$_2$	64	*conv*9$_1$	512 × 512 × 64
*conv*1$_2$	64	*conv*1$_1$	512 × 512 × 64	*conv*9$_1$	64	*merge*9	512 × 512 × 128
*pool*1	2	*conv*1$_2$	512 × 512 × 64	*merge*9	/	*up*9 + *conv*1$_2$	512 × 512 × 128
*conv*2$_1$	128	*pool*1	256 × 256 × 64	*up*9	64	*conv*8$_2$	256 × 256 × 128
*conv*2$_2$	128	*conv*2$_1$	256 × 256 × 128	*conv*8$_2$	128	*conv*8$_1$	256 × 256 × 128
pool$_2$	2	*conv*2$_2$	256 × 256 × 128	*conv*8$_1$	128	*merge*8	256 × 256 × 256
*conv*3$_1$	256	*pool*2	128 × 128 × 128	*merge*8	/	*conv*2$_2$ + *up*8	256 × 256 × 256
*conv*3$_2$	256	*conv*3$_1$	128 × 128 × 256	*up*8	128	*conv*7$_2$	128 × 128 × 256
*pool*3	2	*conv*3$_2$	128 × 128 × 256	*conv*7$_2$	256	*conv*7$_1$	128 × 128 × 256
*conv*4$_1$	512	*pool*3	64 × 64 × 256	*conv*7$_1$	256	*merge*7	128 × 128 × 512
*conv*4$_2$	512	*conv*4$_1$	64 × 64 × 512	*merge*7	/	*conv*3$_2$ + *up*7	128 × 128 × 512
*drop*4	0.5	*conv*4$_2$	64 × 64 × 512	*up*7	256	*conv*6$_2$	64 × 64 × 512
*pool*4	2	*drop*4	64 × 64 × 512	*conv*6$_2$	512	*conv*6$_1$	64 × 64 × 512
*conv*5$_1$	1024	*pool*4	32 × 32 × 512	*conv*6$_1$	512	*merge*6	64 × 64 × 1024
*conv*5$_2$	1024	*conv*5$_1$	32 × 32 × 1024	*merge*6	/	*drop*4 + *up*6	64 × 64 × 1024
*drop*5	0.5	*conv*5$_2$	32 × 32 × 1024	*up*6	512	*drop*5	32 × 32 × 1024

not consider the interactions among different local regions sufficiently. These interactions are very important for the crack region prediction, especially when we only utilize the independent pixel-level classification loss function to train the model. On one hand, interactions among different local regions can avoid the interferences of the outliers. On the other hand, they can decrease the fractions in the predicted crack maps. In these cases, applying a reasonable interaction representation strategy to conventional U-Net is necessary for the crack detection task. Recently, self-attention strategy (Fig. 3) has achieved significant performance for this issue [43], which efficiently models relationships among different separated spatial regions in the image. Because of the aforementioned reasons, we incorporate self-attention layer into conventional U-Net architecture. Considering the computational load, we only add the self-attention layer to the *drop*5 layer, which depicts more general information of the image. Specifically, self-attention strategy is described as follows.

The original feature maps from the previous hidden layer $\mathbf{x} \in \mathbf{R}^{C \times N}$ are first transformed into two feature spaces, f and g, to calculate the attention map, where $f(\mathbf{x}) = \mathbf{W}_f\mathbf{x}$ and $g(\mathbf{x}) = \mathbf{W}_g\mathbf{x}$. The interactions among different local regions can be calculated with Eq. (2).

$$\alpha_{j,i} = \frac{e^{s_{ij}}}{\sum_{i=1}^{N} e^{s_{ij}}}, \tag{2}$$

where $s_{ij} = f(\mathbf{x}_i)^T g(\mathbf{x}_j)$, and $\alpha_{j,i}$ indicates the extent which the model attends to the i_{th} location when synthesizing the j_{th} region. Additionally, the output of the attention layer is $\mathbf{o} = (\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_j, \ldots \mathbf{o}_N) \in \mathbf{R}^{C \times N}$, the element in $\mathbf{o}$ is calculated with Eq. (3).

$$\mathbf{o}_j = \sum_{i=1}^{N} \alpha_{j,i} h(\mathbf{x}_i), \tag{3}$$

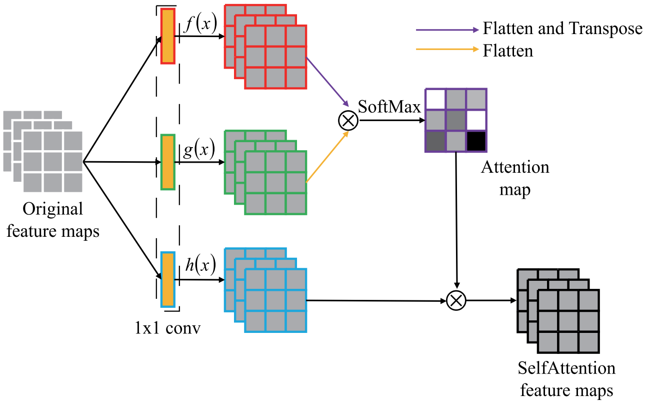


Fig. 3. The Self-attention strategy, which divides the original feature maps into three subbranches, one is to maintain the spatial structure information of the original feature maps, and the other two are used to calculate the attention map. Additionally, ⊗ represents matrix multiplication.

where $h(\mathbf{x}_i) = \mathbf{W}_h\mathbf{x}_i$. In Eqs. (2) and (3), $\mathbf{W}_f \in \mathbf{R}^{\hat{C}\times C}$, $\mathbf{W}_g \in \mathbf{R}^{\hat{C}\times C}$ and $\mathbf{W}_h \in \mathbf{R}^{\hat{C}\times C}$ are three projection matrices, which map original feature maps to three different representation spaces, and they are implemented as three 1 × 1 convolution layers respectively. Specifically, C is 1024, and $\hat{C}$ is set to 256 in this paper.

Besides, like the procedure in [43], we multiply the output of the attention layer by a balance hyperparameter and add back to the input feature. The final output of the self-attention layer is formulated as Eq. (4).

$$\mathbf{y}_i = \gamma\mathbf{o}_i + \mathbf{x}_i, \tag{4}$$

where γ is the factor to balance two terms, which is set to 0.2 in this paper.

3.3.2. Loss function

Besides the network architecture, loss function is also vital for the performance of the detection, because parameters of the network are learned through the loss in a backward propagation way. Because we consider the road crack detection task as a salient object detection one, a pixel-level binary classification loss term is necessary. The conventional cross entropy loss is formulated as Eq. (5),

$$\ell_c = -\frac{1}{WH}\sum_{w=1}^{W}\sum_{h=1}^{H}\left(y^{(w,h)}\log\hat{y}^{(w,h)} + \left(1-y^{(w,h)}\right)\log\left(1-\hat{y}^{(w,h)}\right)\right), \tag{5}$$

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 probability of $(w,\ h)_{th}$ pixel belongs to the crack category respectively. As can be seen, ℓ_c gives same importance weights to different categories, which can train the network well when the sample distribution is balanced. However, the sample distribution of crack data is severely imbalanced (sample scale of crack is much smaller than that of non-crack), the ill-posed classifier will be obtained if directly using the conventional cross entropy loss to train the network. To address this issue, and inspired by the Focal loss [44], we propose a weighted classification loss based on the conventional cross entropy loss function, which avoids the hyperparameter-setting procedure compared to Focal loss. The proposed weighted cross entropy loss is formulated as Eq. (6).

$$\ell_c^w = -\frac{1}{WH}\sum_{w=1}^{W}\sum_{h=1}^{H}\left(\mu\cdot y^{(w,h)}\log\hat{y}^{(w,h)} + (1-\mu)\cdot\left(1-y^{(w,h)}\right)\log\left(1-\hat{y}^{(w,h)}\right)\right), \tag{6}$$

where μ is the importance weight of crack category, which can be calculated through Eq. (7).

$$\mu = \frac{1}{WH}\sum_{w=1}^{W}\sum_{h=1}^{H}\mathrm{I}\{y^{(w,h)} = 0\}, \tag{7}$$

where I{ · } is the indicator function, it equals to 1 if the condition is satisfied and 0 otherwise. ℓ_c^w gives different importance weights to different categories, specifically, importance weights for categories with larger scale are smaller while for smaller ones are larger. This weighted classification loss can alleviate the influence of imbalanced data for model training, and address the ill-posed classifier problem to a ceratin extent.

Besides the weighted cross entropy loss, we design a novel interaction loss to explicitly depict the relationships among different pixels in the image, which is defined as Eq. (8), and the visualized results can be seen in Fig. 4.

$$\ell_i = ||\hat{y}^T\hat{y} - y^Ty||_F^2, \tag{8}$$

where y is the groundtruth of the image, $\hat{y}$ is the predicted saliency map, and $||\cdot||_F$ represents the Frobenius norm. Specifically, $\ell_i^{(w,h)}$ is the $(w,\ h)_{th}$ element of ℓ_i, which represents the difference of relationship of the i_{th} and j_{th} column between predicted saliency map and the groundtruth.

Finally, the overall loss function is consisted of the weighted cross entropy loss and the interaction loss, which is defined as Eq. (9).

$$\ell = \ell_c^w + \lambda^2\cdot\ell_i, \tag{9}$$

where λ is a hyperparameter to balance the relationship between weighted cross entropy loss and interaction loss, which is set to 3×10^{-2} in this paper.

3.4. Data augmentation strategy

This subsection details the proposed data augmentation strategy. Actually, we can directly utilize the aforementioned weighted cross entropy loss to reduce the influence of the imbalanced distributed data, however, which may be too restricted. Specifically, attaching too big importance weight to categories with a very small sample scale is likely to be severely interfered by the outliers, and the training procedure will be not stable. More importantly, even though the training performance is significantly satisfactory, the testing performance is difficult to meet our expectations. In these cases, we propose a groundtruth-expansion based augmentation strategy to address this issue further. Generally speaking, our augmentation strategy is based on the assumption that, pixels in the neighbourhood of real cracks more likely belong to the crack category, compared to others. The detailed procedures are described as follows.

We utilize the truncated expansion strategy to augment the dataset. Specifically, if a pixel belongs to crack category in the original groundtruth, each pixel in its particular neighborhood is set to the crack category in the expanded groundtruth. The neighborhood size is a hyperparameter, which reflects the expansion degree. Some expanded samples are shown in Fig. 5.

Generally speaking, the proposed weighted cross entropy loss and the data augmentation strategy are used to address the data imbalance issue in road crack detection task. Additionally, the self-attention module and the proposed interaction loss term are used to propagate the spatial structure information from raw image to corresponding predicted saliency map well.

4. Experiments

This section details the experiments, including datasets, contrasting methods, experimental settings, measurement metrics and

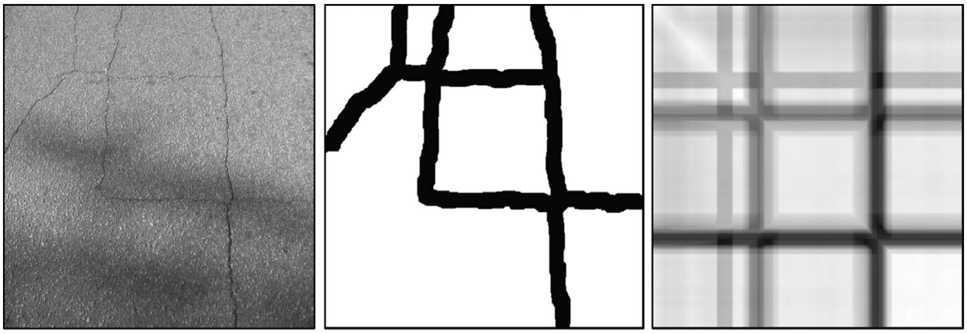


Fig. 4. The schematic diagram of our proposed interaction mechanism, which calculates the interactions among different pixels in the image through considering the 2-order information of the predicted saliency map. The left image is the original raw image, and the middle and the right image are the predicted crack saliency map and its corresponding 2-order interaction map respectively.

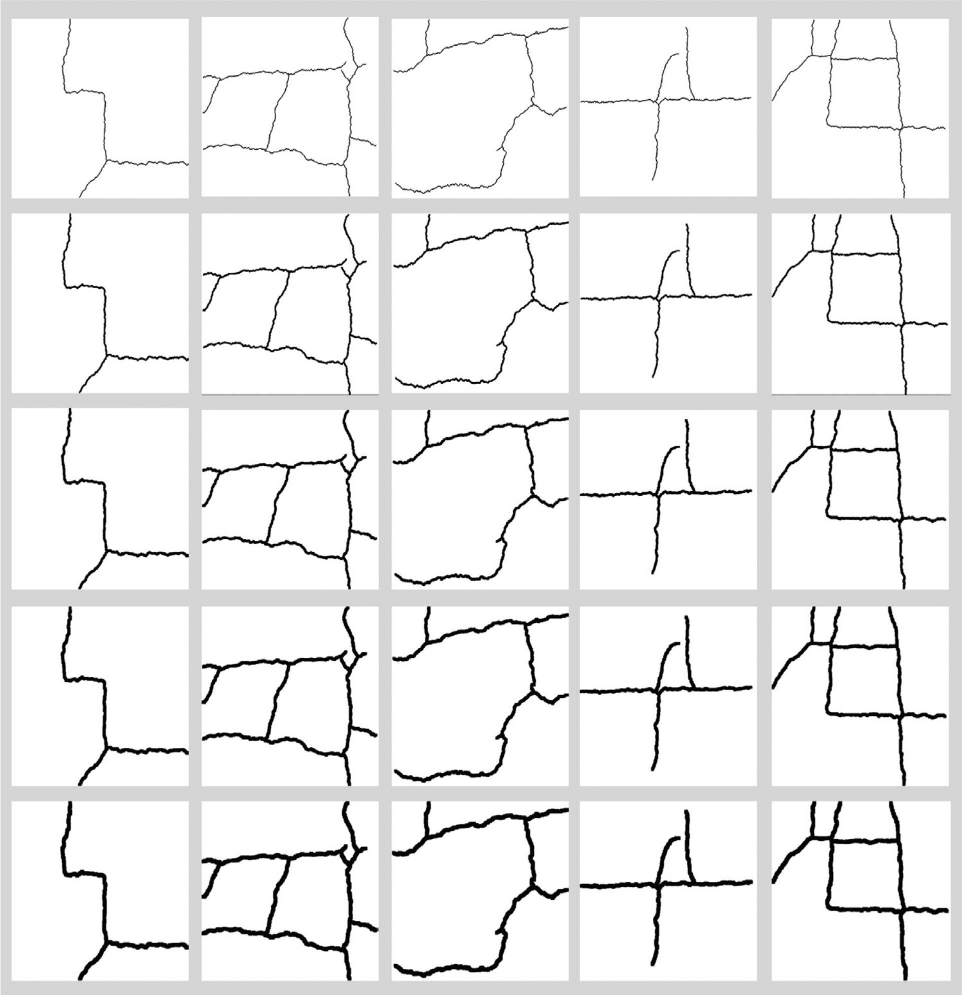


Fig. 5. The visualized results of truncated expansion strategy. Images in the first row are the original groundtruths, images in the second, third, fourth and fifth rows represents the corresponding expanded groundtruths, and the neighborhood sizes are 3, 5, 7, 9, respectively.

experimental results, which are described in Sections 4.1, 4.2, 4.3, 4.4 and 4.5, respectively.

4.1. Datasets

This subsection introduces three datasets used in this paper, CrackTree200 dataset [45], ALE dataset [46] and CrackForest dataset [47].

CrackTree200 dataset [45] contains 206 pavement images with fixed size of 800 × 600, whose crack types are various. This dataset is challenging due to the complex inference factors such as shadows, occlusions and low contrast. Some samples are shown on the first and second row in Fig. 8.

ALE dataset [46] contains three sub-datasets, Aigle-RN, ESAR and LCMS. Aigle-RN contains 38 images with pixel-level annotations. ESAR contains 15 fully annotated crack images, which is

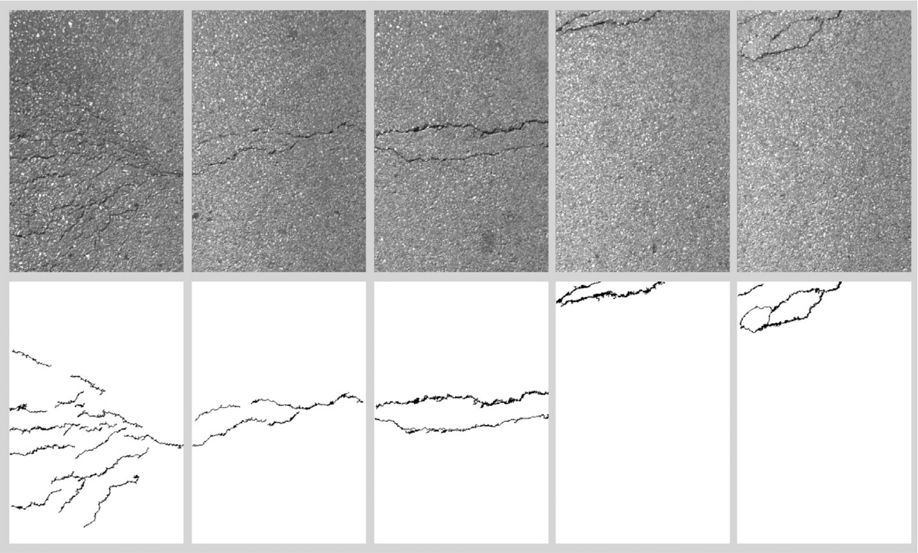


Fig. 6. Five samples of the ALE dataset, the first and second row respectively represents the original road images and corresponding groundtruth.

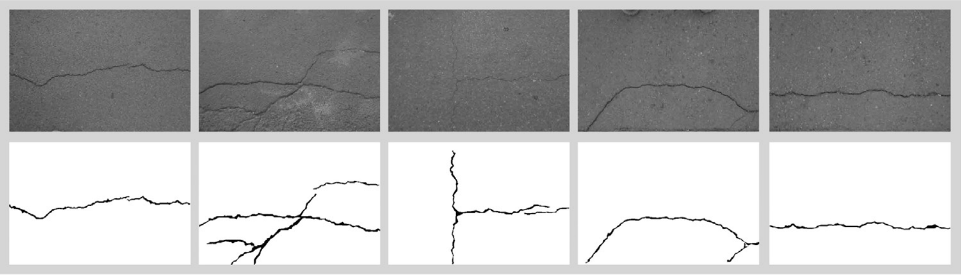


Fig. 7. Five samples of the CrackForest dataset, the first and second row respectively represents the original road images and corresponding groundtruth.

acquired through a static acquisition system with no controlled lighting. LCMS contains 5 pixel-level annotated crack images. Some samples are shown in Fig. 6.

CrackForest dataset [47] includes 118 images, which are photographed in Beijing, China. Each image has hand labeled ground truth, and images in this dataset contain noises such as shadows, oil spots and water stains. Some samples are shown in Fig. 7.

4.2. Contrasting methods

This subsection introduces five contrasting methods, which contain two patch classification based ones and three salient object detection based ones. The patch classification based contrasting methods include deep convolutional neural networks with transfer learning (DCNNTL) [48] and deep learning-based crack detection using convolutional neural network with Naive Bayes data fusion (NB-CNN). The three salient object detection based contrasting methods include fully convolutional networks (FCNs) [49], dilated convolutional network (DilatedCN) [50], and U-Net [51].

DCNNTL [48] transfers to apply deep convolutional networks trained on ImageNet to automatically detect cracks in pavement images.

NB-CNN [52] combines a convolutional networks and a Naive Bayes data fusion scheme to enhance the crack detection system.

FCN [49] is the ground-breaking work for semantic image segmentation task with deep neural network, which realizes the segmentation in an end-to-end way and achieves competitive performances.

DilatedCN [50] expands the receptive filed through adding holes with different sizes to the convolution filters step by step, but not shrinks the feature maps layer by layer, which can maintain more structure information compared to conventional convolutional networks.

U-Net [51] propagates more structure, contexture and other detailed information from the original raw image to the predicted semantic map through skip-connection strategy, which can obtain finer prediction mask.

4.3. Experimental settings

In this paper, for each supervised based contrasting method, we choose 50% images of the dataset to train the network, and leave the others to be the testing set. For the deep neural network (DNN) based method, we train the model through mini-batch SGD with batch size 5. The initial learning rate is set to 1.0×10^{-3}, the momentum is set to 0.9 and each model is trained for 50 epoches. Additionally, the DNN based methods are implemented with Keras, and the testing platform is X99UD4 of GIGABYTE, GPU (8G × 8) of Titan X.

4.4. Measurement metrics

In this paper, we utilize three commonly used metrics on salient object detection task to measure the effectiveness of the proposed method, including *precision (P)*, *recall (R)* and *f-measure* (F_β).

$$P = \frac{|M \cap G|}{|M|}, \tag{10}$$

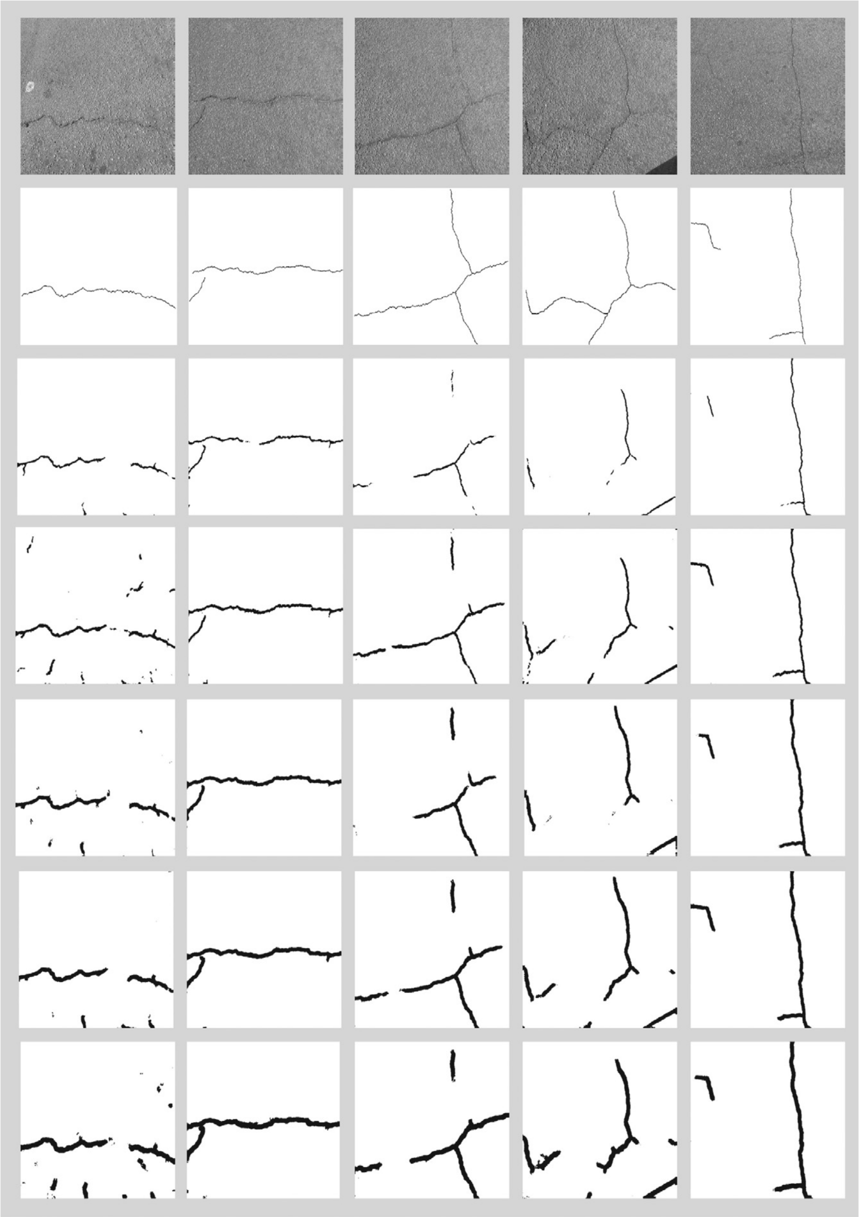


Fig. 8. Masks predicted using models which trained by augmented groundtruths with different hyperparameter k. The first row represents the original road image and the second row represents the corresponding groundtruth. Additionally, the third to seventh row represent masks predicted using model which trained by augmented groundtruth with hyperparameter $k = 1, 3, 5, 7, 9,$ respectively.

$$R = \frac{|M \cap G|}{|G|}, \tag{11}$$

where M is the binary mask transformed from saliency map with a specific threshold, and G is the corresponding groundtruth. Usually, both P and R can not evaluate the quality of a saliency map comprehensively. In this case, F-measure is proposed as a weighted harmonic mean of P and R with a non-negative weight β, which is defined as Eq. (12).

$$F_\beta = \frac{\left(1 + \beta^2\right) \cdot P \cdot R}{\beta^2 \cdot P + R}, \tag{12}$$

where β is a hyperparameter, which is used to balance the P and R. As is suggested in [53,54], β^2 is set to 0.3 to increase the importance of the P value. The reason for weighting precision more than recall is that the recall rate is not as important as precision.

4.5. Experimental results

This subsection reports the experimental results on three datasets, including ablation experiment and comparison experiments on three datasets. Specifically, because the main contributions of this paper contain three components, since a series of ablation experiments are applied to verify the their effectiveness. Additionally, to verify the superiority of the proposed method, we compare our method with other existing methods on three datasets. The ablation and comparison experiments are described in Section 4.5.1 and Section 4.5.2, respectively.

4.5.1. Ablation experiments

The contributions of the proposed method mainly contains five components, image preprocessing strategy, self-attention strategy, weighted cross entropy loss, interaction loss and data augmentation strategy. To verify each contribution promoting the

Table 2
Experimental results with different hyperparameters.

k	1	3	5	7	9
P (%)	**29.07**	18.91	13.01	11.56	9.24
R (%)	85.66	90.13	90.12	**90.80**	90.65
F_β (%)	**30.75**	20.23	14.00	12.64	9.98

Table 3
Ablation experimental results.

/	Pre-P	Self-A	I-L	P (%)	R (%)	F_β (%)
Ablation	–	–	–	18.91	90.13	20.23
	√	–	–	19.49	91.27	20.84
	√	√	–	19.84	92.05	21.21
	√	√	√	**20.96**	**93.44**	**22.39**

crack detection performance, we do the ablation experiments on Crack200 dataset [45]. For the convenience of the subsequent ablation experiments, before the ablation experiments, an important hyperparameter for data augmentation, neighborhood size (k) should be set at first. In this case, we use U-Net + Data Augmentation + Weighted Cross Entropy Loss (UDAWCEL), with size k, to choose the best hyperparameter k. The experimental results are shown in Table 2, and some samples of the visualized results are shown in Fig. 8.

According to the experimental results from Table 2 and the visualized results from Fig. 8, we set the hyperparameter $k = 3$ for CrackTree200 dataset. Specifically, when k is larger than 3, recall (R) remains unchanged essentially while precision (P) decreases rapidly. In these cases, we set $k = 3$ as the optimal hyperparameter for CrackTree200 dataset.

After the hyperparameter k is set, we do the ablation experiments about other four contributions. The experimental results are shown in Table 3, where Pre-P, Self-A and I-L denote image preprocessing strategy, self-attention strategy and interaction loss strategy respectively. Additionally, we use three metrics, precision (P), recall rate (R) and F-measure (F_β), to measure the experimental performances.

From Table 3, we can see that each component of the proposed method contributes to the performance. Specifically, compared with the aforementioned UDAWCEL, it obtains a 2.05% improvement in terms of P and a 2.16% improvement in terms of F_β when applying the proposed preprocessing strategy to it. Additionally, the self-attention module and the proposed interaction loss can also enhance the detection capability and robustness of the proposed method. To depict these points clearly, we visualize some corresponding samples that before and after apply the proposed operation series, which are shown in Fig. 9.

From Fig. 9 we can see that, the proposed operation series improve detection performance to a certain extent. Firstly, the procedures can effectively avoid the noise influence, which can be found in the first column. The reason is that, the preprocessing strategy utilizes multi-scale medfilter and weighted fusion mechanism. As is known, medfilter can reduce the influence of impulse noise, and the multi-scale fusion strategy enhances this point more effectively. Secondly, the proposed operation series improve the crack connectivity of the predicted mask, and samples in the second and fourth columns reflect this point clearly. Additionally, the proposed operation series reduce the missed detection rate and false detection rate, we can find these points clearly in the third and fifth column respectively. Especially, in the fifth column, the proposed operation series address the shadow influence to the predicted mask. These two aforementioned phenomenons are benefited from the spatial structure relationships among different local regions in the original road image and predicted mask, which is realized through the self-attention module and the interaction loss. Specifically, conventional U-Net predicted the category label of each pixel independently, this results in a problem that the predicted mask is easily influenced by the inferences such as noise and shadow. The proposed operation series, self-attention module and interaction loss, consider the spatial structure information of image in terms of region-level and pixel-level respectively, which enhances the robustness of the U-Net through incorporating the 2-order interaction information into the model.

Table 4
Experimental results on three datasets.

Dataset	Method	P (%)	R (%)	F_β (%)
CrackTree200 [45]	DCNNTL [48]	3.83	75.62	4.91
	NB-CNN [52]	9.17	77.00	11.50
	FCN8s [49]	17.73	86.45	18.98
	DilatedCN [50]	24.18	85.32	25.70
	U-Net [51]	**29.07**	85.66	**30.75**
	Ours	20.96	**93.44**	22.39
ALE [46]	DCNNTL [48]	2.23	91.98	2.87
	NB-CNN [52]	3.97	88.51	5.10
	FCN8s [49]	19.30	95.86	20.66
	DilatedCN [50]	20.61	82.29	21.97
	U-Net [51]	**37.33**	95.78	**39.31**
	Ours	25.74	**96.15**	27.40
CrackForest [47]	DCNNTL [48]	9.33	73.10	11.68
	NB-CNN [52]	18.53	78.48	22.49
	FCN8s [49]	**45.61**	77.00	**47.20**
	DilatedCN [50]	42.98	80.56	44.70
	U-Net [51]	40.43	81.46	42.18
	Ours	40.44	**83.78**	42.24

4.5.2. Comparison experiments

This section reports the comparison experiments, the results are shown in Table 4. It is necessary to note that, all the contrasting methods are trained with our weighted cross entropy loss, if not, each of them can not even detect any crack regions for any dataset. From Table 4 we can see that, the proposed method achieves relatively satisfactory performances on three testing datasets, especially on the CrackTree200 dataset, the most challenging one. Specifically, for CrackTree200 dataset, the proposed method obtains 7.78% recall increment compared to the conventional U-Net, the basic model of our proposed method. Then, we note that, the precision rate of our method is lower than that of U-Net's. The main reason is that, we utilize the expanded groundtruth to train the network and further improve the detection precision, and the detected cracks become wider while the recall improves. Actually, we can use a corresponding corrosion strategy to improve the precision, which is not detailed here since it is not the main contribution of our method. As we can see, the experimental results on ALE and CrackForest datasets have shown the similar rule. Additionally, we find an interesting phenomenon that, for ALE dataset, FCN8s achieves better performance than Dialted Network. The main reason is that, compared to CrackTree200 and CrackForest datasets, crack distribution in ALE dataset is dense. In other words, too large receptive field may bring to strong inferences among them and influence the detection performance further. Finally, salient object based road crack detection methods (FCN8s, DilatedCN, U-Net, and Ours) have achieved much more competitive performances by comparison with those of patch classification based ones (DCNNTL and NB-CNN) in general, which demonstrates the rationality of using salient object detectors to achieve road crack detection.

In summary, for an arbitrary fully convolutional network, (1) the self-attention mechanism and the proposed interaction loss can effectively improve its spatial structure representation capability, (2) the proposed truncated expansion strategy and weighted classification loss can alleviate the severe sample imbalance problem, and these two points can improve the detection performance to a large extent. Besides road crack detection task, the proposed framework can address other complex structural prediction

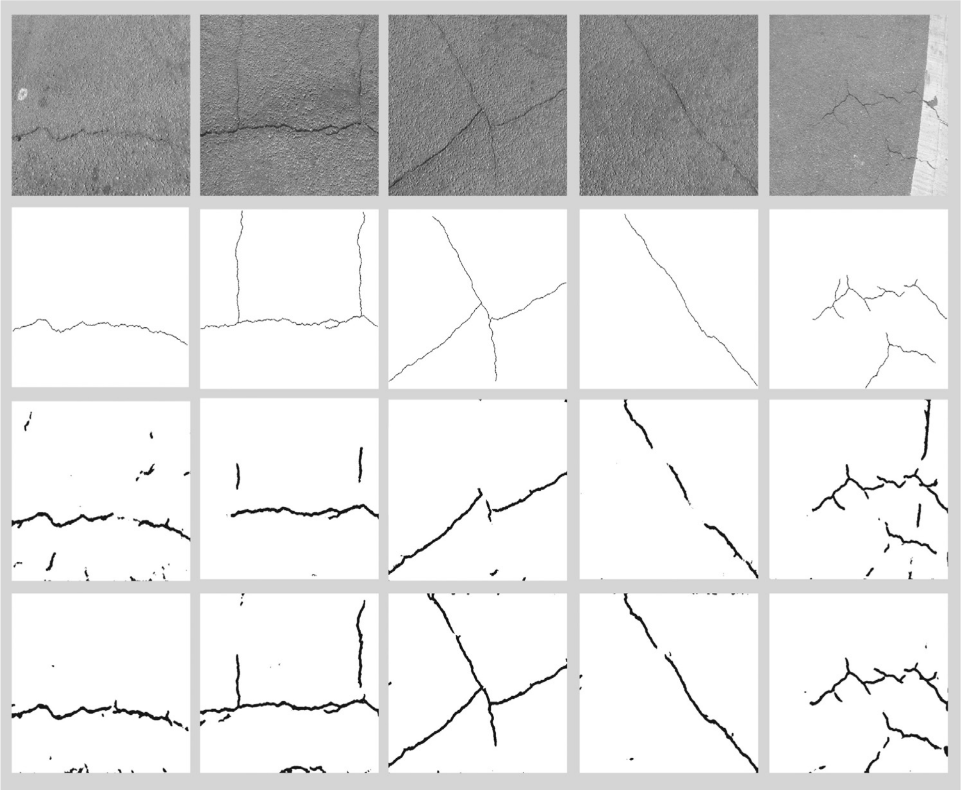


Fig. 9. Some corresponding samples that before and after apply the proposed operation series to UDAWCEL, the first row represents the original road image, the second, third and fourth row represents the groundtruth, the results of UDAWCEL, and the results after applying the proposed operation series to UDAWCEL.

tasks with severe sample imbalance problems well. However, the proposed method still does not break through the framework of supervised learning, which limits its practical application especially when huge gaps exist between training and testing samples. In the future, we will combine our method with some weakly supervised learning methods, semi supervised learning methods, and unsupervised learning methods to address the aforementioned problem.

5. Conclusion and future work

In this paper, we propose a distribution equalization learning mechanism for road crack detection. Specifically, we consider the road crack detection task as a salient object detection one, and obtain the crack saliency map directly from the designed network. Firstly, we design a median-filtering based preprocessing strategy to avoid the influences from noises and other interferences. Secondly, we propose an explicit augmentation strategy and a novel weighted cross entropy loss term to address the severe data imbalance problem in road crack detection task. Thirdly, we incorporate the proposed interaction loss term and the popular self-attention module into our network to propagate the detailed spatial structure information of image to corresponding predicted saliency map well. Finally, experimental results on three public and challenging datasets demonstrate the effectiveness and robustness of the proposed method.

Even though the performance of the proposed method is in accordance with our expectation relatively, the predicted saliency results are a little coarse because of the truncated expansion augmented strategy. In the future, we will try to incorporate the reasonable operations, such as "corrosion", into the network to obtain more accurate crack saliency map. Additionally, the proposed weighted cross entropy loss in this paper is based on the sample number information only, which ignores their structure relationships, and we will consider this point further.

Author Contribution

J. Fang designed the algorithm, implemented the program and wrote the paper. B. Qu and Y. Yuan polished the paper.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported in part by the National Science Fund for Distinguished Young Scholars under Grant 61825603, in part by the State Key Program of National Natural Science Foundation of China under Grant 61632018, in part by the National Key R&D Program of China under Grant 2017YFB0502900, in part by the National Natural Science Foundation of China under Grant 61925112, Grant 61806193, 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 CAS "Light of West China" Program under Grant XAB2017B15.

References

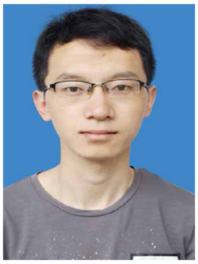
[1] Y. Xia, B. Chen, S. Weng, Y.-Q. Ni, Y.-L. Xu, Temperature effect on vibration properties of civil structures: a literature review and case studies, J. Civil Struct. Health Monitor. 2 (1) (2012) 29–46.

[2] P. Cornwell, C.R. Farrar, S.W. Doebling, H. Sohn, Environmental variability of modal properties, Exp. Tech. 23 (6) (1999) 45–48.

[3] J.G. Chen, N. Wadhwa, Y.-J. Cha, F. Durand, W.T. Freeman, O. Buyukozturk, Modal identification of simple structures with high-speed video using motion magnification, J. Sound Vib. 345 (2015) 58–71.

[4] Y.-J. Cha, J. Chen, O. Büyüköztürk, Output-only computer vision based damage detection using phase-based optical flow and unscented Kalman filters, Eng. Struct. 132 (2017) 300–313.

[5] I. Abdel-Qader, O. Abudayyeh, M.E. Kelly, Analysis of edge-detection techniques for crack identification in bridges, J. Comput. Civil Eng. 17 (4) (2003) 255–263.
[6] D. Ziou, S. Tabbone, et al., Edge detection techniques-an overview, Pattern Recognit. Image Anal. C/C of Raspoznavaniye Obrazov I Analiz Izobrazhenii 8 (1998) 537–559.
[7] Y. Yuan, J. Fang, X. Lu, Y. Feng, Remote sensing image scene classification using rearranged local features, IEEE Trans. Geosci. Remote Sens. 57 (3) (2019) 1779–1792.
[8] C. Wang, X. Bai, S. Wang, J. Zhou, P. Ren, Multiscale visual attention networks for object detection in VHR remote sensing images, IEEE Geosci. Remote Sens. Lett. 16 (2) (2019) 310–314, doi:10.1109/LGRS.2018.2872355.
[9] J. Fang, X. Cao, GAN and DCN based multi-step supervised learning for image semantic segmentation, in: Proceedings of the First Chinese Conference Pattern Recognition and Computer Vision - PRCV 2018, Guangzhou, China, November 23–26, 2018, Proceedings, Part II, 2018, pp. 28–40, doi:10.1007/978-3-030-03335-4_3.
[10] Y.-J. Cha, W. Choi, O. Büyüköztürk, Deep learning-based crack damage detection using convolutional neural networks, Comput.-Aid. Civil Infrastruct. Eng. 32 (5) (2017) 361–378.
[11] S. Teidj, A. Khamlichi, A. Driouach, Identification of beam cracks by solution of an inverse problem, Procedia Technol. 22 (2016) 86–93.
[12] Rabinovich, Givoli, Vigdergauz, Xfem-based crack detection scheme using a genetic algorithm, Int. J. Numer. Methods Eng. 71 (9) (2010) 1051–1080.
[13] C.M. Yeum, S.J. Dyke, Vision-based automated crack detection for bridge inspection, Comput.-Aid. Civil Infrastruct. Eng. 30 (10) (2015) 759–770.
[14] L.I. Rudin, S. Osher, E. Fatemi, Nonlinear total variation based noise removal algorithms, in: Proceedings of the Eleventh International Conference of the Center for Nonlinear Studies on Experimental Mathematics: Computational Issues in Nonlinear Science: Computational Issues in Nonlinear Science, 1992.
[15] K. Zhang, H. Cheng, A novel pavement crack detection approach using pre-selection based on transfer learning, in: Proceedings of the International Conference on Image and Graphics, Springer, 2017, pp. 273–283.
[16] F.-C. Chen, M.R. Jahanshahi, NB-CNN: deep learning-based crack detection using convolutional neural network and Naïve Bayes data fusion, IEEE Trans. Ind. Electron. 65 (5) (2018) 4392–4400.
[17] F. Yang, L. Zhang, S. Yu, D. Prokhorov, X. Mei, H. Ling, Feature pyramid and hierarchical boosting network for pavement crack detection, arXiv:1901.06340 (2019).
[18] A.F. Frangi, W.J. Niessen, R.M. Hoogeveen, V. Walsum T, M.A. Viergever, Model-based quantitation of 3-D magnetic resonance angiographic images, IEEE Trans. Med. Imaging 18 (10) (2002) 946–956.
[19] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Trans. Pattern Anal. Mach. Intell. 20 (11) (1998) 1254–1259.
[20] K. Fu, C. Gong, I.Y.-H. Gu, J. Yang, Normalized cut-based saliency detection by adaptive multi-level region merging, IEEE Trans. Image Process. 24 (12) (2015) 5671–5683.
[21] C. Gong, D. Tao, W. Liu, S.J. Maybank, M. Fang, K. Fu, J. Yang, Saliency propagation from simple to difficult, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2531–2539.
[22] X. Zhou, Z. Liu, C. Gong, W. Liu, Improving video saliency detection via localized estimation and spatiotemporal refinement, IEEE Trans. Multimed. 20 (11) (2018) 2993–3007.
[23] Y. Xie, H. Lu, Visual saliency detection based on bayesian model, in: Proceedings of the 2011 18th IEEE International Conference on Image Processing, IEEE, 2011, pp. 645–648.
[24] M.M. Cheng, N.J. Mitra, X. Huang, P.H.S. Torr, S.M. Hu, Global contrast based salient region detection, IEEE Trans. Pattern Anal. Mach. Intell. 37 (3) (2015) 569–582.
[25] W. Zhu, S. Liang, Y. Wei, J. Sun, Saliency optimization from robust background detection, in: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23–28, 2014, 2014, pp. 2814–2821, doi:10.1109/CVPR.2014.360.
[26] C. Yang, L. Zhang, H. Lu, X. Ruan, M. Yang, Saliency detection via graph-based manifold ranking, in: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23–28, 2013, 2013, pp. 3166–3173, doi:10.1109/CVPR.2013.407.
[27] B. Jiang, L. Zhang, H. Lu, C. Yang, M.H. Yang, Saliency detection via absorbing Markov chain, in: Proceedings of the IEEE International Conference on Computer Vision, 2013.
[28] J. Han, D. Zhang, X. Hu, G. Lei, J. Ren, W. Feng, Background prior based salient object detection via deep reconstruction residual, IEEE Trans. Circuits Syst. Video Technol. 25 (8) (2015) 1309–1321.
[29] H. Liu, R. Wang, S. Shan, X. Chen, Learning multifunctional binary codes for both category and attribute oriented retrieval tasks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3901–3910.
[30] X. Peng, J. Feng, S. Xiao, W.-Y. Yau, J.T. Zhou, S. Yang, Structured autoencoders for subspace clustering, IEEE Trans. Image Process. 27 (10) (2018) 5076–5086.
[31] Z. Huang, H. Zhu, J.T. Zhou, X. Peng, Multiple marginal fisher analysis, IEEE Transactions on Industrial Electronics (2018).
[32] J. Wang, H. Jiang, Z. Yuan, M.M. Cheng, X. Hu, N. Zheng, Salient object detection: a discriminative regional feature integration approach, Int. J. Comput. Vis. 123 (2) (2017) 251–268.
[33] L. Wang, H. Lu, X. Ruan, M. Yang, Deep networks for saliency detection via local estimation and global search, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7–12, 2015, 2015, pp. 3183–3192, doi:10.1109/CVPR.2015.7298938.
[34] R. Zhao, W. Ouyang, H. Li, X. Wang, Saliency detection by multi-context deep learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7–12, 2015, 2015, pp. 1265–1274, doi:10.1109/CVPR.2015.7298731.
[35] G. Lee, Y. Tai, J. Kim, Deep saliency with encoded low level distance map and high level features, in: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27–30, 2016, 2016, pp. 660–668, doi:10.1109/CVPR.2016.78.
[36] Y. Liu, R. Wang, S. Shan, X. Chen, Structure inference net: object detection using scene-level context and instance-level relationships, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6985–6994.
[37] S.H. Khan, M. Hayat, M. Bennamoun, F.A. Sohel, R. Togneri, Cost-sensitive learning of deep feature representations from imbalanced data, IEEE Trans. Neural Netw. Learn. Syst. 29 (8) (2018) 3573–3587.
[38] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, Smote: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (1) (2002) 321–357.
[39] K. Li, X. Kong, L. Zhi, W. Liu, J. Yin, Boosting weighted elm for imbalanced learning, Neurocomputing 128 (5) (2014) 15–21.
[40] M. Kukar, I. Kononenko, Cost-sensitive learning with neural networks, in: Proceedings of the ECAI, 1998, pp. 445–449.
[41] C. Yuan, H.T. Lin, S.W. Yang, Cost-aware pre-training for multiclass cost-sensitive deep learning, Comput. Sci. (2015).
[42] V. Raj, S. Magg, S. Wermter, Towards effective classification of imbalanced data with convolutional neural networks, in: Proceedings of the Artificial Neural Networks in Pattern Recognition - 7th IAPR TC3 Workshop, ANNPR 2016, Ulm, Germany, September 28–30, 2016, Proceedings, 2016, pp. 150–162, doi:10.1007/978-3-319-46182-3_13.
[43] H. Zhang, I. Goodfellow, D. Metaxas, A. Odena, Self-attention generative adversarial networks, arXiv:1805.08318 (2018).
[44] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2980–2988.
[45] Q. Zou, Y. Cao, Q. Li, Q. Mao, S. Wang, Cracktree: automatic crack detection from pavement images, Pattern Recognit. Lett. 33 (3) (2012) 227–238.
[46] R. Amhaz, S. Chambon, J. Idier, V. Baltazart, Automatic crack detection on two-dimensional pavement images: an algorithm based on minimal path selection, IEEE Trans. Intell. Transp. Syst. 17 (10) (2016) 2718–2729.
[47] Y. Shi, L. Cui, Z. Qi, F. Meng, Z. Chen, Automatic road crack detection using random structured forests, IEEE Trans. Intell. Transp. Syst. 17 (12) (2016) 3434–3445.
[48] K. Gopalakrishnan, S.K. Khaitan, A. Choudhary, A. Agrawal, Deep convolutional neural networks with transfer learning for computer vision-based data-driven pavement distress detection, Construct. Build. Mater. 157 (2017) 322–330.
[49] E. Shelhamer, J. Long, T. Darrell, Fully convolutional networks for semantic segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (4) (2017) 640–651.
[50] Y. Bengio, Y. LeCun (Eds.), Proceedings of the 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2–4, 2016, Conference Track Proceedings, 2016 https://iclr.cc/archive/www/doku.php%3Fid=iclr2016:accepted-main.html.
[51] O. Ronneberger, P. Fischer, T. Brox, U-net: convolutional networks for biomedical image segmentation, in: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 234–241.
[52] F.-C. Chen, M.R. Jahanshahi, NB-CNN: deep learning-based crack detection using convolutional neural network and Naïve Bayes data fusion, IEEE Trans. Ind. Electron. 65 (5) (2017) 4392–4400.
[53] R. Achanta, S. Hemami, F. Estrada, S. Süsstrunk, Frequency-tuned salient region detection, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2009) CONF, 2009, pp. 1597–1604.
[54] F. Perazzi, P. Krähenbühl, Y. Pritch, A. Hornung, Saliency filters: Contrast based filtering for salient region detection, in: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 733–740.



Jie Fang Received B.S. degree in school of electronic engineering from XiDian University, Xi'an 710126, Shaanxi, P. R. 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 710119, Shaanxi, P. R. China and with the University of Chinese Academy of Sciences, Beijing 100049, P. R. China. His research interests include Artificial intelligence, Machine Learning and Image Understanding.

Bo Qu is with the Key Laboratory of Spectral Imaging Technology CAS, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, Shaanxi, P. R. China. His research interests include Image Processing, Artificial Intelligence and Machine Learning.

Yuan Yuan (M'05-SM'09) is currently a Full Professor with the Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, P. R. China. Her current research interests include Visual Information Processing and Image/Video Content Analysis.