External Attention Based TransUNet and Label Expansion Strategy for Crack Detection

Jie Fang, Chen Yang, Yuetian Shi, Nan Wang, and Yang Zhao

***Abstract*— Crack detection is an indispensable premise of road maintenance, which can provide early warning information for many road damages and save repair costs to a large extent. Because of the security and convenience, many image processing technique (IPT) based crack detection methods have been proposed, but their performances often cannot meet the requirements of practical applications because of the complex texture structure and seriously imbalanced categories. To address the aforementioned problem, we present an external attention based TransUNet for crack detection. Specifically, we tackle the TransUNet as the backbone of our detection framework, which can propagate the detailed texture information from shallow layers to corresponding deep layers through skip connections. Besides, the Transformer Block equipped in the second last convolution layer of the encoding component can explicitly model the long-range dependency of different regions in an image, which improves the structural representation ability of the framework and hence alleviates the interference from shadow, noise, and other negative factors. In addition, the External Attention Block equipped in the last convolution layer of the encoding component can effectively exploit the dependency of crack regions among different images, and further enhance the robustness of the framework. Finally, combined with the Focal Loss, the proposed label expansion strategy can further alleviate the category imbalance problem through transforming semantic categories of non-crack pixels distributed in the neighbors of corresponding crack pixels.**

***Index Terms*— Crack detection, TransUNet, external attention, label expansion.**

I. Introduction

ROAD maintenance is one of the most important duties of traffic units, while timely detection and early warning are the premise [1]. Most of the existing crack detection methods belong to the scope of field investigation, which usually accompany several limitations such as personal risk and hysteresis [2]. Recently, because of its security and convenience, the image processing technique (IPT) based methods especially the deep learning based ones [3], [4] have achieved relatively competitive performances. Even though, since the complex structure and seriously imbalanced categories, the existing methods cannot dig out the latent projection dependencies among road images and corresponding category maps, which means the requirements of practical applications cannot be satisfied.

To address the aforementioned issues, we propose an external attention based TransUNet [5] for crack detection, which can improve the detection performance through enhancing the structural representation capability of the model and decreasing quantity differences between crack and non-crack categories. The flowchart of the proposed method is shown in Fig. 1. Specifically, the skip connections can propagate the details from shallow layers to corresponding deep ones, and further increase the detection performance. Besides, the Transformer Block equipped in the second last convolution layer of the encoding component can explicitly model the long-range dependencies among different local regions in the same image, which enhances the structural representation capability of the model and further alleviate the influences from interference information such as shadow and noise. In addition, the External Attention Block [6] equipped in the last convolution layer of the encoding component can effectively exploit the dependencies among interesting regions in different images, which improves the adaptivity and robustness of the framework for different practical scenes. Finally, we present a label expansion strategy and combine it with the Focal Loss [7] based on Binary Cross Entropy, which can address the missed and false detection problem caused by seriously category imbalance.

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

1) We consider crack detection as a pixel-wise classification task, and utilize the TransUNet as the backbone network to exploit the latent projection relationship among road images and corresponding crack maps.
2) We incorporate the External Attention Block into the last convolution layer of the encoding component to enhance the representation capability of the model for dependency of interesting regions among different images.
3) We present a label expansion strategy to increase the amount of crack samples and combine it with the Binary Cross Entropy based Focal Loss to alleviate the seriously category imbalance problem.

Manuscript received August 4, 2021; revised November 10, 2021 and January 3, 2022; accepted February 22, 2022. The Associate Editor for this article was Z. Duric. *(Corresponding author: Jie Fang.)*

Jie Fang is with the School of Telecommunication and Information Engineering, Xi'an University of Posts and Telecommunications, Xi'an, Shaanxi 710121, China, and also with Corporation of Shaanxi Wukong Clouds Information and Technology, Xi'an, Shaanxi 710000, China (e-mail: 2443952262@qq.com).

Chen Yang is with the Ministry of Science and Technology, Pudong Development Bank, Xi'an, Shaanxi 710065, China.

Yuetian Shi and Nan Wang are with the Key Laboratory of Spectral Imaging Technology CAS, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, Shaanxi 710119, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China.

Yang Zhao is with the College of Transportation Engineering, Chang'an University, Xi'an, Shaanxi 710064, China.

Digital Object Identifier 10.1109/TITS.2022.3154407

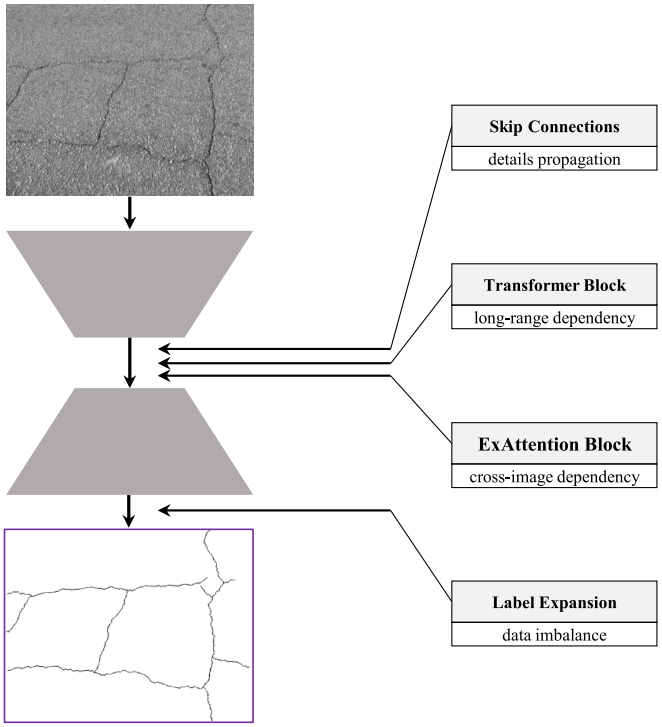

Fig. 1. The flowchart of the proposed method, which utilizes an encoder-decoder based network to finalize the projection among road images and corresponding crack maps. Furthermore, the skip connections, transformer block, external attention block, and label expansion strategy are incorporated into the framework to enhance the representation capability, adaptivity, and robustness of the model.

The remainder of the paper is organized as follows. Sec. II introduces some existing salient object detection methods. Sec. III describes the proposed framework in brief. Sec. IV reports the experimental results, and Sec. V concludes the paper.

II. Related Works

This Sec. introduces the existing IPT based crack detection approaches, and divides them into handcrafted feature based methods and deep learning based ones according to the feature types they used.

A. Handcrafted Feature Based Methods

At the very beginning, it is assumed that crack regions in an image can be distinguished easily by statistical handcrafted features. Delagnes and Barba [8] proposed a Markov random field based rectilinear structure extraction algorithm for crack detection, besides the algorithm itself, it still needs subsequent post-processing to obtain the crack regions in an image. Cord and Chambon [9] proposed a supervised crack detection algorithm based on Adaboost, which enhances its representation capability through sufficiently exploiting the texture patterns of road images. Oliveira and Correia [10] introduced an assignment methodology of crack severity levels, which computes the estimate according to the width of detected crack. Shi *et al.* [11] proposed a random structured forest based crack detection method, which uses the integral channel features to redefine the tokens that constitute a crack, and further obtains more robust representations for cracks with intensity inhomogeneity. Tang and Gu [12] proposed a hybrid crack detection and segmentation method, which first utilizes histogram based threshold algorithm to roughly locate the cracks, and then uses the mathematics morphology and B-spline based snake model to refine the locations of cracks. Oliveira and Correia [13] proposed an entropy and image dynamic threshold strategy based crack detection method, which first uses a dynamic thresholding to search potential crack pixels, and then utilizes a classification system to find out the real crack pixels.

B. Deep Learning Based Methods

Even though the handcrafted feature based crack detection methods have achieved gradually improvements, their performances still cannot meet the requirements of the practical applications because it is difficult for hand-crafted features to depict the latent projection relationships among road images and corresponding category maps. With the rapid development of hardware and software in artificial intelligence field, because of the strong nonlinear representation capability of deep neural networks, deep learning based methods have achieved competitive performances on many computer vision tasks such as scene recognition [14], [15], image retrieval [16], semantic segmentation [17], [18] visual attention prediction [19], [20], image-audio translation [21], [22], crowd counting and localization [23], [24] and salient object detection [25], [26].

As for crack detection, Oliveira and Correia [27] proposed an image processing toolbox and characterization, which contribute to the development of this field to a large extent. Gopalakrishnan *et al.* [28] divided the road image into different patches, and transferred a deep convolutional neural network trained on ImageNet to predict the category of each patch. Zhang *et al.* [29] proposed to use convolutional neural network to infer the corresponding road saliency map from original optical image, which also use the block-breaking then classify strategy to finalize the detection. Chen and Jahanshahi [30] proposed a convolutional neural network and naïve Bayes data fusion based crack detection framework, which enhances the overall performance and robustness of the model through aggregating the information extracted from each individual frame. Zou *et al.* [31] proposed to learn hierarchical convolutional features to improve the detection performance, which fuses feature maps of shallow and deep layers from conventional convolutional networks to obtain representation with both spatial structure details and discriminative semantic attributes. Wu *et al.* [3] proposed a sample and structure-guided network for crack detection, which uses the U-Net as the backbone of the detection network, then incorporates self-attention module into its bottle-neck to enhance its representation capability and utilizes the conventional Focal Loss to optimize the network. Similar to the S^2GNet in [3], Fang *et al.* [4] proposed a distribution equalization learning mechanism for crack detection, which utilizes an auxiliary interaction loss term to enhance the relationships among different local regions in an image. Yang *et al.* [32]

proposed feature pyramid and hierarchical boosting network for pavement crack detection, which improves the detection performance from feature fusion and decision fusion. Liu *et al.* [33] proposed a two-step crack detection method based on YOLOv3 object detection model and U-Net segmentation model. Kim *et al.* [34] proposed a LeNet-5 based shallow convolutional network for crack detection, which can save computational load and annotation costs to a large extent because of its lightweight model architecture. Even though, since it is based on patch-classification strategy and hence cannot directly give the crack map of whole road image, this method is time-consuming.

III. Proposed Method

This Sec. details the proposed crack detection algorithm. Specifically, Sec. III-A introduces the overview diagram. Sec. III-B gives the multiscale filtering fusion based image pre-processing strategy. Sec. III-C and Sec. III-D demonstrates the inner processing flow of Transformer Block and External Attention Block respectively. Sec. III-E introduces the proposed label expansion strategy, and Sec. III-F gives the cost function of the proposed algorithm.

A. Overview

The overview diagram of the proposed crack detection method is shown as Fig. 2. Firstly, the input road image is fed into an multiscale filtering fusion based pre-processing module, which can filter out the noises from the original image. Then the fused image is fed into an external attention based TransUNet to obtain its corresponding binary crack map. Compared with the existing TransUNet, we shift the transformer block from the last encode layer to the penultimate layer, and incorporate the external attention module into the last encode layer to improve the structural representation capability of the model. Specifically, the transformer block is used to depict the interactions of different local region in an image, while the external attention module is used to dig out the latent dependency relationships of interesting regions among different images. The location relationship between transformer block and external attention module is decided by the semantic levels, the dependencies among different local regions are lower than those among different images and hence we fed feature maps from relatively shallow layers into it. In addition, the proposed label expansion strategy and Binary Cross Entropy based Focal Loss are combined to address the model collapse issue caused by the seriously category imbalance between crack and non-crack categories.

B. Multiscale Filtering Fusion

To filter out influences of shadows among different particles and other noises, we design a multiscale filtering fusion strategy to process the input images, which is defined as Equation 1.

$$\mathbf{I}_F = \sum_{n=1}^{N} \delta_n \mathrm{F}_{(n)}^{M}(\mathbf{I}) \tag{1}$$

where N denotes the number of mean filters, which is set to 4 in this paper. $\mathbf{I}$ denotes the input image. $\mathrm{F}_{(n)}^{M}(\cdot)$ denotes the n_{th} mean filter with size $(2^n-1)\times(2^n-1)$. δ_n denotes the weight coefficient of n_{th} filter, which can be calculated by Equation 2.

$$\delta_n = \frac{e^{-n}}{\sum_{k=1}^{N} e^{-k}} \tag{2}$$

It can be seen from Equation 2, we give relatively bigger weights to filters with smaller size, which is used to remain the detailed information of the crack regions in the image.

C. Transformer Block

Considering the curvilinear distribution rather than regional distribution particularity of crack in the image, different from [5], we incorporate the transformer block (TransBlock) into the second last convolution layer but not the last convolution layer of the encoder component, which can model the detailed dependencies among different local regions in an image better. To introduce the flow process of the proposed framework, we demonstrate the inner architecture of TransBlock here in brief, and its diagram is shown in Fig. 3.

The input feature cube is first divided into a series of patches in spatial dimension, and these patches are vectorized as x_p. Then a trainable linear projection layer is used to map x_p into a latent D-dimensional feature space. In addition, learnable position features are added to the patch features to maintain position information and further encode the patch position information, which is defined as Equation 3.

$$z_0 = \left[x_p^1\phi; x_p^2\phi; \cdots; x_p^N\phi\right] + \phi_{pos} \tag{3}$$

where x_p^1 to x_p^N denote the vectorized patches, and N denotes the number of patches. $\phi \in \mathcal{R}^{(P^2\cdot C)\times D}$ denotes the patch projection, P and C denote the size and channel number of aforementioned feature patch respectively. $\phi_{pos} \in \mathcal{R}^{N\times D}$ denotes the position feature.

The Transformer Block contains K=12 Transformer layers, and each layer consists of a Multihead Self-Attention module and a Multi-Layer Perception module. In these cases, the output of k_{th} Transformer layer can be obtained by Equation 4.

$$\begin{aligned} \hat{z}_k &= \mathrm{MSA}\left(\mathrm{LN}\left(z_{k-1}\right)\right) + z_{k-1}, \\ z_k &= \mathrm{MLP}\left(\mathrm{LN}\left(\hat{z}_k\right)\right) + \hat{z}_k, \end{aligned} \tag{4}$$

where $\mathrm{LN}(\cdot)$ denotes the layer normalization. MSA and MLP respectively denotes the Multihead Self-Attention [35] and Multi-Layer Perception modules. In addition, the final representation z_K is reshaped to feature cube which has the same spatial size as the input feature cube. To illustrate the inner architecture of the network more clearly, we review the Multihead Self-Attention (MSA) [35] here. Multihead Self-Attention projects the queries, keys and values several times with different learnable linear projections, then performs attention function on each of these projected groups in parallel, and finally concatenates the results together as the output.

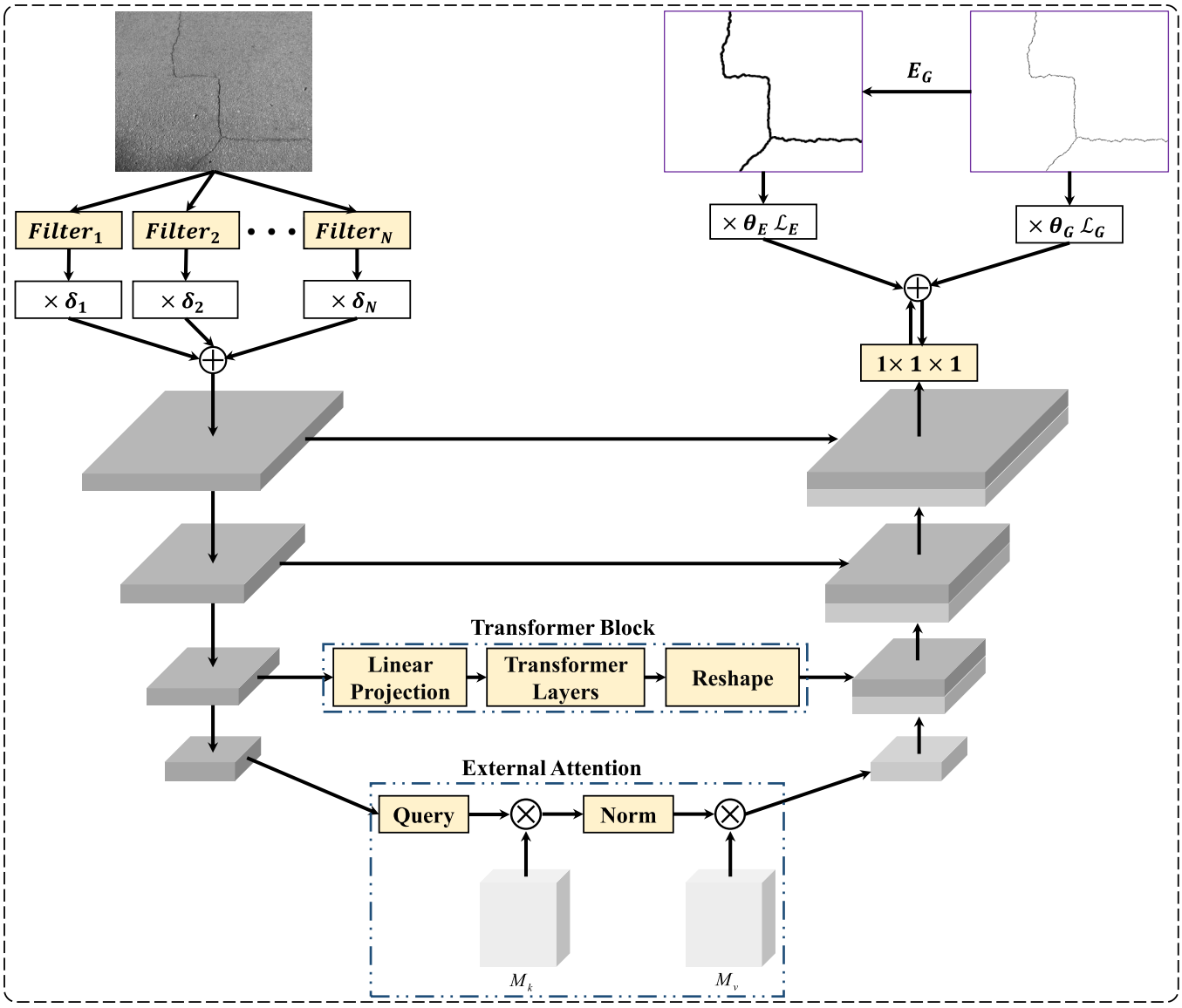

Fig. 2. The proposed crack detection framework, which mainly contains a multiscale filtering fusion based pre-prosessing strategy, an external attention based TransUNet and a label expansion mechanism based optimization strategy. Specifically, the multiscale filtering fusion based pre-processing strategy can alleviate the interferences from particle shadows and other noises. Besides, the external attention based TransUNet cannot only explicitly model the dependencies among different local regions in an image, but also model the dependencies of interesting regions cross different images. In addition, the label expansion strategy can alleviate the model collapse situation through reasonably increasing the sample of crack category.

The formulation is defined as Equation 5,

$$\begin{aligned} &MSA\ (Q, K, V) = Cat\left(h_1, h_2, \ldots, h_{N_h}\right) W^O \\ &s.t.\ h_i = Att\left(QW_i^Q, KW_i^K, VW_i^V\right) \end{aligned} \tag{5}$$

where Q, K, V is the query, key, and value of the input group respectively. W^O, W^Q, W^K, and W^V denotes the projection matrices of output, query, key, and value respectively. $Att\,(\cdot)$ denotes the Scaled Dot-Production attention function, which is defined as Equation 6,

$$Att\,(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V, \tag{6}$$

where d_k denotes the dimension of Q and K.

D. External Attention Mechanism

Even though the Transformer Block aforementioned in Sec. III-C can enhance the dependencies among different local regions in an image, it still cannot model the dependencies of interesting regions among different images. However, this cross-image dependency is of importance for the structural representation capability and robustness of the model because similar categories regions should have similar latent semantic information even among different images. In these cases, we incorporate the novel External Attention module into the last convolution layer of the encoder component, which uses two learnable memory units $M_k \in \mathcal{R}^{d\times S}$ and $M_v \in \mathcal{R}^{d\times S}$ as key and value to improve the capability of the network. It is noteworthy that M_k and M_v are independent to the input, which act memories for all samples in the training set.

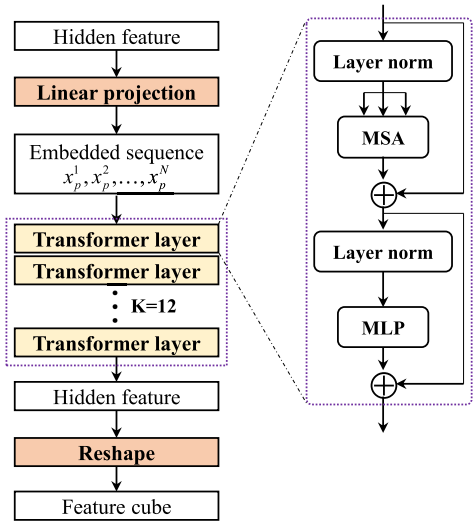

Fig. 3. The inner architecture of the Transformer Block, which mainly contains a linear projection layer, 12 transformer layers, and a reshape layer. Specifically, each transformer layer contains a multi-head self-attention module, a multi-layer perception module, and two layer norm modules.

Specifically, the flow process of External Attention module is defined from Equation 7 to Equation 8 in turn.

$$A = (\alpha)_{n,m} = \text{Norm}\left(F_{in} M_k^T\right), \tag{7}$$

$$F_{out} = AM_v, \tag{8}$$

where F_{in} denotes the input feature, F_{out} denotes the output feature. Norm $(\cdot)$ represents the double-normalization operator [36], and its flow process is sequentially defined from Equation 9 to Equation 11.

$$(\tilde{\alpha})_{n,m} = F_{in} M_k^T \tag{9}$$

$$\hat{\alpha}_{n,m} = \frac{\exp\left(\tilde{\alpha}_{n,m}\right)}{\sum_t \exp\left(\tilde{\alpha}_{t,m}\right)} \tag{10}$$

$$\alpha_{n,m} = \frac{\hat{\alpha}_{n,m}}{\sum_t \hat{\alpha}_{n,t}} \tag{11}$$

where $\alpha_{n,m}$ denotes the $(n,m)_{th}$ element in attention map A.

E. Label Expansion Strategy

Even though the Focal loss [7] can alleviate the category imbalance problem for some common situations, it cannot address the problem in crack detection field well. Investigate its reasons, under the situation of seriously category imbalance, directly apply weighted balance strategy may lead to the over-sensitivity of the model for category with small scale and further result in the increment of false positive. To address the aforementioned problem, we present a label expansion strategy to increase the amount of crack samples, which is based on the assumption that, if a pixel belongs to crack category, pixels distributed in its neighborhood are more likely to crack category, compared to other pixels. In this case, we transform the semantic category of non-crack pixels distributed in the neighborhood of crack ones from non-crack to crack category, and the formulation is defined as Equation 12.

$$\begin{aligned} &\mathbf{S}_E^{\left(w-\left[\frac{M}{2}\right]:w+\left[\frac{M}{2}\right],h-\left[\frac{M}{2}\right]:h+\left[\frac{M}{2}\right]\right)} = \mathbf{I}_0^{M\times M}, \\ &s.t.\ \mathbf{S}^{(w,h)} = 0 \end{aligned} \tag{12}$$

where $\mathbf{S}$ denotes the true category map. $\mathbf{S}_E$ denotes the expanded category map. M denotes the size of expansion window. $[\cdot]$ denotes the Down Integral Function. $\mathbf{I}_0^{M\times M}$ denotes a $M \times M$ matrix whose elements are all zeros. It can bee seen from Equation 12, if a pixel in $\mathbf{S}$ belongs to crack category, pixels of its $M \times M$ neighbourhood in $\mathbf{S}_E$ are labeled to crack category.

F. Optimization Function

The optimization function of the proposed algorithm is defined as Equation 13, which mainly contains two terms.

$$\mathcal{L} = \theta_G \mathcal{L}_G\left(\hat{\mathbf{S}}, \mathbf{S}\right) + \theta_E \mathcal{L}_E\left(\hat{\mathbf{S}}, \mathbf{S}_E\right), \tag{13}$$

where $\hat{\mathbf{S}}$ denotes the predicted category map. $\mathcal{L}_G$ denotes the original classification loss while $\mathcal{L}_E$ denotes the label expansion based classification loss, which are defined as Equation 14 and Equation 15 respectively. θ_G and θ_E are two hyperparameters to balance the relative importance of $\mathcal{L}_G$ and $\mathcal{L}_E$.

$$\begin{aligned} \mathcal{L}_G\left(\hat{y}, y\right) = &-\frac{1}{WH}\sum_{w=1}^{W}\sum_{h=1}^{H}\mu_G y_{w,h}\left(1-\hat{y}_{w,h}\right)^{\gamma_G}\log\hat{y}_{w,h} \\ &+\left(1-\mu_G\right)\left(1-y_{w,h}\right)\left(\hat{y}_{w,h}\right)^{\gamma_G}\log\left(1-\hat{y}_{w,h}\right), \end{aligned} \tag{14}$$

$$\begin{aligned} \mathcal{L}_E\left(\hat{y}, y\right) = &-\frac{1}{WH}\sum_{w=1}^{W}\sum_{h=1}^{H}\mu_E y_{w,h}\left(1-\hat{y}_{w,h}\right)^{\gamma_E}\log\hat{y}_{w,h} \\ &+\left(1-\mu_E\right)\left(1-y_{w,h}\right)\left(\hat{y}_{w,h}\right)^{\gamma_E}\log\left(1-\hat{y}_{w,h}\right), \end{aligned} \tag{15}$$

where W and H denotes the width and height of the image respectively. $y_{w,h}$ and $\hat{y}_{w,h}$ denotes the label and predicted score of $(w,h)_{th}$ pixel in the image. μ_G, γ_G, μ_E, and γ_E are four hyperparameters to adjust the relative importance of different categories in $\mathcal{L}_G$ and $\mathcal{L}_E$. Specifically, according the suggestion from [7], we set $\gamma_G = 2$, $\gamma_E = 2$. In addition, we use category frequency to give values of μ_G and μ_E, which are defined as Equation 16 and Equation 17 respectively.

$$\mu_G = \frac{NWH}{\sum_{n=1}^{N}\sum_{w=1}^{W}\sum_{h=1}^{H}\mathrm{I}\left\{y_{n,w,h}^G = 1\right\}}, \tag{16}$$

$$\mu_E = \frac{NWH}{\sum_{n=1}^{N}\sum_{w=1}^{W}\sum_{h=1}^{H}\mathrm{I}\left\{y_{n,w,h}^E = 1\right\}}, \tag{17}$$

where $\mathrm{I}\{\cdot\}$ denotes the indictor function, it equals to 1 when the condition satisfies and 0 others. N denotes the number of images in the training set. $y_{n,w,h}^G$ and $y_{n,w,h}^E$ denote the $(w,h)_{th}$ element of original label map y^G and expanded label map y^E respectively.

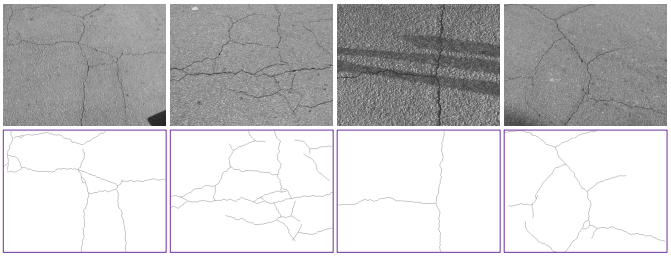

Fig. 4. Some examples of Cractree200 dataset.

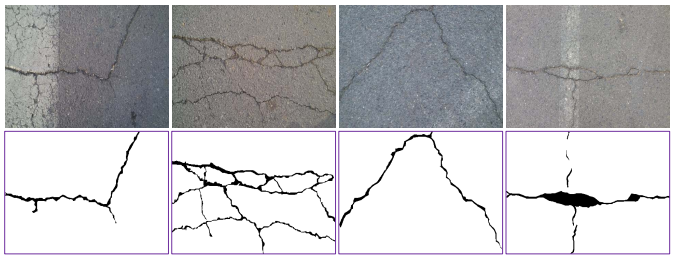

Fig. 5. Some examples of CrackForest dataset.

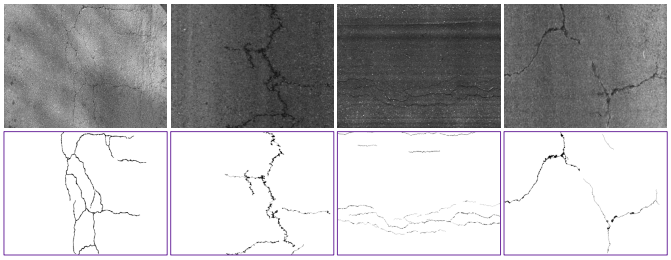

Fig. 6. Some examples of ALE dataset.

Fig. 7. Some examples of CrackPV dataset.

IV. Experiments

This Sec. reports the experiments, including datasets and evaluation metrics, contrasting methods, experimental settings, hyper-parameter analysis and experimental results.

A. Datasets and Evaluation Metrics

1) Datasets: In order to validate the effectiveness of the proposed algorithm, we conduct the experiments on four datasets, including Cracktree200 dataset [37], Crack Forest dataset [38], ALE dataset [11], and CrackPV dataset [3].

Cracktree200 dataset contains 206 images with corresponding pixel-level annotations, which is a challenging one because of shadows, occlusions, and other interference factors. Some samples are shown in Fig. 4.

Crack Forest dataset contains 118 images with fixed size of 480 × 320, and each sample in this dataset has its hand-annotated crack label map. Some samples of this dataset are shown in Fig. 5.

ALE dataset contains 58 images from three subset, including Aigle-RN, LCMS, and ESAR. Images in each subset have their own unique apparent characteristics, and which increase the detection difficulty. Some samples are shown in Fig. 6.

CrackPV dataset contains more than 500 road images photographed in practical scenes, which are interfered by a series of controllable and uncontrollable interference factors such as exposure intensity and weather condition and further improve its difficulty. It is noteworthy that images in this dataset do not have corresponding pixel-level annotations, but which can still used as an auxiliary data to validate the performance of different methods in practical applications through visualized results. Some samples of CrackPV dataset are shown in Fig. 7.

2) Evaluation Metrics: Three common metrics in salient object detection field are used to measure the performance of different crack detection algorithms, including precision (P), recall (R), and F-measure (F_β). Specifically, P and R are defined in Equation 18.

$$P = \frac{TP}{TP + FP},\ R = \frac{TP}{TP+FN}, \tag{18}$$

where TP denotes the number of samples whose true label and predicted label are both positive. FP denotes the number of samples with negative true label and positive predicted label. FN denotes the number of samples with positive true label and negative predicted label. In addition, F_β is a weighted harmonic mean of P and R, which is defined as Equation 19.

$$F_\beta = \frac{\left(1+\beta^2\right)\cdot P \cdot R}{\left(\beta^2 \cdot P\right) + R}, \tag{19}$$

where β is a hyperparameter to balance the relative importance between P and R. Detailedly, β^2 is set to 0.3 to emphasize the precision according to the suggestion in [39].

B. Contrasting Methods

To verify the superiority of the proposed method, we compare it with some state-of-the-art ones, including deep neural network with transfer learning (DCNNTL) [28], deep learning crack detection using convolutional neural network with Naive Bayes data fusion (NB-CNN) [30], U-type Network (UNet) [40], sample and structure-guided network (S^2GNet) [3], distribution equalization learning mechanism (DELM) [4], and TransUNet [5].

C. Experimental Settings

All of the algorithms are implemented with Pytorch, and the testing platform is GeForece RTX 2080. The proposed model is trained by mini-batch Adam strategy with batch size 2. The weight decay is set to 5×10^{-5}. The initial learning rate is set to 5×10^{-4}, the momentum is set to 0.9 and each of the contrasting models is trained for 200 epoches. In addition, we first utilize the large scale of salient object detection dataset DUTs [41] as an auxiliary one to train the network to enhance its structural representation capability, and

TABLE I
ABLATION EXPERIMENTAL RESULTS ON CRACKTREE200 DATASET

Transformer Block	External Attention Block	Label Expansion	Multiscale filtering fusion	$P(\%)$	$R(\%)$	$F_\beta(\%)$
–	–	–	–	16.36	57.22	19.59
√	–	–	–	27.62	31.89	28.50
√	√	–	–	31.26	59.44	35.10
√	√	√	–	37.54	49.33	39.73
√	√	√	√	**39.79**	**47.67**	**41.37**

TABLE II
THE PERFORMANCES WITH DIFFERENT TRANSFORMER LAYERS ON CRACKTREE200 DATASET

N_T	$P(\%)$	$R(\%)$	$F_\beta(\%)$
4	29.97	39.85	31.79
8	35.31	61.23	39.13
12	**39.91**	52.72	**42.28**
16	39.19	**53.72**	41.80

TABLE III
THE PERFORMANCES WITH DIFFERENT ($\frac{\theta_G}{\theta_E}$) RATIOS ON CRACKTREE200 DATASET

$\frac{\theta_G}{\theta_E}$	$P(\%)$	$R(\%)$	$F_\beta(\%)$
1	17.02	**86.03**	20.88
4	21.39	82.69	25.80
16	29.78	73.73	34.53
64	36.03	64.30	40.10
128	38.41	56.18	41.43
256	39.68	56.58	42.63
512	**40.72**	54.55	**43.26**
1024	40.55	53.25	42.91

then use the specific crack dataset to fine-tune the model to capture detailed information of road images.

D. Hyper-Parameter Analysis

To choose the most suitable hyperparametric combination of the number of transformer layers N_T and ratio between θ_G and θ_E, we have conducted the experiments on Cracktree200 dataset, the results are shown in Table II and Table III respectively.

From Table II we can see that F_β shows a trend of first growth and then decline according to the increment of N_T under the situation of $\frac{\theta_G}{\theta_E} = 256$, which reaches the peak when $N_T = 12$. Investigate its reason, the spatial structural of representation can be enhanced by the increment of N_T within a certain extent since the stronger region-relationship mining capacity, while too big N_T may bring overfitting, spatial dispersion and other interference factors.

From Table III we can see that F_β increases first and then decreases with the increment of $\frac{\theta_G}{\theta_E} = 256$ under the situation of $N_T = 12$, which has peaked at 42.28% when $\frac{\theta_G}{\theta_E} = 512$. The reason is that too big $\frac{\theta_G}{\theta_E}$ cannot balance the samples of different categories, while too small $\frac{\theta_G}{\theta_E}$ results in the overemphasation of pseudo expanded label.

E. Experimental Results

1) Ablation Experiments: To verify the effectiveness of each of the proposed component, we conduct the ablation experiments on Cracktree200 dataset. The quantitative results are shown in Table. I, from which we can see that each of the proposed component can contribute to the detection performance. Specifically, the Transformer Block gains a 11.26% increment in terms of Precision because which enhances the interaction relationships and dependencies of different local regions in the image. Besides, External Attention Block gains a 27.55% increment in terms of Recall because which encodes the latent intrinsic similarity information of crack regions among across different images into the model. In addition, Label Expansion strategy gains 4.63% increment in terms of F_β because which alleviates the over sensitivity issues of Focal Loss to crack samples. Finally, multiscale filtering fusion strategy gains 2.25% Precision increment and 1.56 F_β increment respectively, this is because which can avoid the interferences of uncontrolled noises to a large extent.

2) Contrasting Experiments on Three Public Datasets: To validate the superiority of the proposed algorithm, we conduct the experiments on Cracktree200, Crack Forest, and ALE dataset. The quantitative results are shown in Table IV

From Table IV we can see that some common phenomenons of different contrasting methods on all of the three datasets. Firstly, pixel-wise classification based methods have shown better detection results than those of patch-wise classification based methods. Specifically, the performances of DCNNTL and NB-CNN are not as satisfactory as the remainder ones. The reason is that, patch-wise classification based methods are coarse estimation ones, which can only give the category (either "crack" or "non-crack") to each patch of road images, but cannot infer the specific location, width, and other detailed distribution information of cracks in the patch, hence the pixel-wise evaluation metrics are not satisfactory as expected. Secondly, as for pixel-wise classification based methods, S^2GNet and DELM achieve better performances than the traditional UNet, which are benefitted by their special network architectures and optimization styles. Specifically, the incorporations of attention mechanisms of S^2GNet and DELM

TABLE IV
CONTRASTING EXPERIMENTAL RESULTS ON CRACKTREE200 DATASET, CRACK FOREST DATASET, AND ALE DATASET

Dataset	Method	$P(\%)$	$R(\%)$	$F_\beta(\%)$
Cracktree200 [37]	DCNNTL [28]	3.83	75.62	4.91
	NB-CNN [30]	9.17	77.00	11.50
	UNet [40]	14.36	57.22	19.59
	S^2GNet [3]	15.47	90.11	19.03
	DELM [4]	20.96	**93.44**	22.39
	TransUNet [5]	25.93	47.55	28.97
	Ours	**39.79**	47.67	**41.37**
Crack Forest [38]	DCNNTL [28]	9.33	73.10	11.68
	NB-CNN [30]	18.53	78.48	22.49
	UNet [40]	40.43	81.46	42.18
	S^2GNet [3]	43.30	76.23	48.09
	DELM [4]	40.44	**83.78**	42.24
	TransUNet [5]	41.86	75.93	46.69
	Ours	**55.26**	68.14	**57.78**
ALE [11]	DCNNTL [28]	2.23	91.98	2.87
	NB-CNN [30]	3.97	88.51	5.10
	UNet [40]	34.89	94.66	40.84
	S^2GNet [3]	41.88	93.99	48.03
	DELM [4]	25.74	**96.15**	27.40
	TransUNet [5]	49.15	60.52	51.38
	Ours	**53.12**	56.45	**53.84**

contribute to their feature representation capabilities to a large extent, and the enhancements for category with small scale of samples in cost functions improve their sensitivities for "crack" category. 3) TransNet has achieved more significant detection performances than S^2GNet and DELM, because the TransBlock can encode the long-range dependencies among different local regions in an image to corresponding structural representation, which can avoid the influences of shadow and other interference factors. 4) The results of the proposed algorithm surpass those of TransUNet especially on Crack Forest dataset. Investigate its reasons, the multiscale filtering fusion strategy alleviates the influences of controlled and uncontrolled noisy factors, the external attention block strengthens the dependencies of interesting regions among different images, and the label expansion strategy based optimization function balances the sensitivity and robustness of the model for samples with "crack" category. According to the aforementioned analysis, the proposed crack detection algorithm is of effectiveness and superiority.

3) Experiments on CrackPV Dataset: Besides ablation experiments and contrasting experiments, we applied the proposed algorithm on CrackPV dataset to test its practical values. However, there existing two challenges: 1) Samples in CrackPV dataset do not have corresponding crack maps, hence which cannot used to train the model. 2) Lane-lines of images in CrackPV dataset are apparent, which affects the detection performances seriously. In these cases, we present a transfer learning and image addition technique based data processing mechanism to address the aforementioned issues, the procedures are shown in Algorithm 1.

Algorithm 1 Transfer Learning and Image Addition Technique Based Data Processing Mechanism

Input: Imagery series $\{B_k\}_{k=1}^K$ from Crack Forest dataset, imagery series $\{D_k\}_{k=1}^K$ from CrackPV dataset
Output: Fused imagery series $\{BD_k\}_{k=1}^K$.
Processing processes:
1) For each k $(1 \le k \le K)$, using image normalization strategy defined in Equation 20 to obtain the normalized images B_k^N and D_k^N.
2) For each k $(1 \le k \le K)$, using image addition procedure defined in Equation 23. to obtain the normalized fused image BD_k.

$$B^N_{k,(w,h)} = \frac{B_{k,(w,h)} - M_{B_k}}{\sigma_{B_k}},$$
$$D^N_{k,(w,h)} = \frac{D_{k,(w,h)} - M_{D_k}}{\sigma_{D_k}}, \tag{20}$$

where k denotes the index of the imagery. $B^N_{k,(w,h)}$ and $D^N_{k,(w,h)}$ denotes the $(w,h)_{th}$ element of normalized B_k^N and the $(w,h)_{th}$ element of normalized D_k^N respectively. M_{B_k} and M_{D_k} denotes the mean value of B_k and D_k respectively, which are defined in Equation 21. σ_{B_k} and σ_{D_k} denotes the standard deviation of B_k and D_k respectively, which are defined in Equation 22.

$$M_{B_k} = \frac{1}{WH}\sum_{w=1}^{W}\sum_{h=1}^{H} B_{k,(w,h)}$$
$$M_{D_k} = \frac{1}{WH}\sum_{w=1}^{W}\sum_{h=1}^{H} D_{k,(w,h)}, \tag{21}$$
$$\sigma_{B_k} = \sqrt{\frac{1}{WH-1}\left(B_{k,(w,h)} - M_{B_k}\right)^2}$$
$$\sigma_{D_k} = \sqrt{\frac{1}{WH-1}\left(D_{k,(w,h)} - M_{D_k}\right)^2}, \tag{22}$$

where W and H denote the width and height of the image respectively.

$$BD_k = \psi B_k + (1-\psi)D_k, \tag{23}$$

where ψ denotes a hyperparameter distributed in the open interval of zero to one, which is used to balance the relative importance of B_k and D_k. After the above procedures, we used $\{BD_k, S_k\}_{k=1}^K$ series to optimize the network, in which S_k denotes the crack map correspond to B_k.

Supplemental Instructions The original intention of the proposed image addition technique is to alleviate the differences between training set Crack Forest and practical application set

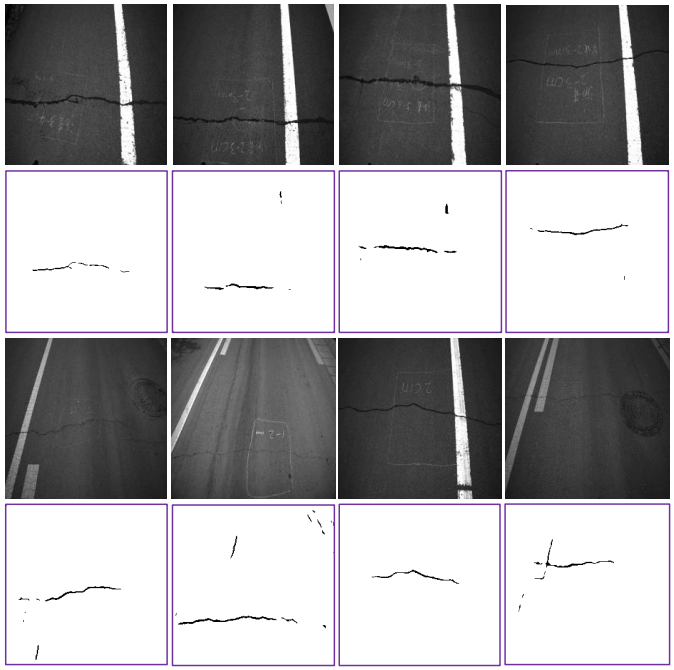

Fig. 8. Some visualized examples of CrackPV dataset.

CrackPV. Actually, when fusing two images that both have cracks, this strategy only penalizes one crack as foreground and which may degrade the performance of the network. In this case, we use the hyperparameter ψ in Equation 23 to control the interference degree of additional images from CrackPV set in training phase. Specifically, relatively bigger ψ can reduce impact of crack samples from CrackPV set, while relatively smaller ψ can alleviate the domain gap of training set and testing set. In other words, the proposed image addition technique can enhance the performance of the network by setting appropriate hyperparameter ψ according to the attribute information of image pairs from Crack Forest and CrackPV set, and we set $\psi = 0.8$ in this paper.

Some visualized examples of our method on CrackPV dataset are shown in Fig. 8, from which we can see that the proposed method can meet the practical scenes even if we only use the transfer learning strategy to optimize the model. Specifically, the proposed model can alleviate the influences from various factors such as lane-line, illumination changes, etc. The main reasons include the strong feature representation capability of the model since its skip connection, Transformer block, and external attention mechanisms, and 2) the incorporation of intrinsic characteristics of target domain to source domain in training phase through the image addition strategy enhances the adaptivity of the model for practical scenes.

V. Conclusion

In this paper, we present an external attention based TransUNet and label expansion strategy for crack detection, which enhances the existing model from the perspectives of both network architecture and optimization mechanism. Specifically, the transformer block and external attention block can respectively explicitly model long-range dependencies among different local regions in an image and dependencies of interesting regions in different images, the former can improve the structural representation capability of the model and hence alleviate influences from different interference factors, while the latter can ensure the inconsistence of crack regions among different images and further increase the adaptivity and robustness of the model for different practical scenes. In addition, the proposed label expansion strategy can increase the sample amount of crack categories through transforming the category properties of pixels distributed in the neighbourhood of crack pixels in real label map, and combine the Binary Cross Entropy based Focal Loss to alleviate the model collapse problem because of seriously category imbalance. Finally, the experimental results on three public and one practical photographed datasets validate the effectiveness and superiority of the proposed method.

References

[1] N. Said *et al.*, "Natural disasters detection in social media and satellite imagery: A survey," *Multimedia Tools Appl.*, vol. 78, no. 22, pp. 31267–31302, Nov. 2019.

[2] I. Abdel-Qader, O. Abudayyeh, and M. E. Kelly, "Analysis of edge-detection techniques for crack identification in bridges," *J. Comput. Civil Eng.*, vol. 17, no. 4, pp. 255–263, Oct. 2003.

[3] S. Wu, J. Fang, X. Zheng, and X. Li, "Sample and structure-guided network for road crack detection," *IEEE Access*, vol. 7, pp. 130032–130043, 2019.

[4] J. Fang, B. Qu, and Y. Yuan, "Distribution equalization learning mechanism for road crack detection," *Neurocomputing*, vol. 424, pp. 193–204, Feb. 2021.

[5] J. Chen *et al.*, "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.

[6] M.-H. Guo, Z.-N. Liu, T.-J. Mu, and S.-M. Hu, "Beyond self-attention: External attention using two linear layers for visual tasks," 2021, *arXiv:2105.02358*.

[7] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] X. Lu *et al.*, "Jm-net and cluster-SVM for aerial scene classification," in *Proc. IJCAI*, 2017, pp. 2386–2392.

[15] J. Fang, X. Cao, P. Han, and D. Wang, "Multidimensional attention learning for VHR remote sensing imagery recognition," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[16] Y. Chen and X. Lu, "Deep category-level and regularized hashing with global semantic similarity learning," *IEEE Trans. Cybern.*, vol. 51, no. 12, pp. 6240–6252, Dec. 2021.

[17] T. Lei, R. Wang, Y. Zhang, Y. Wan, C. Liu, and A. K. Nandi, "DefED-Net: Deformable encoder-decoder network for liver and liver tumor segmentation," *IEEE Trans. Radiat. Plasma Med. Sci.*, vol. 6, no. 1, pp. 68–78, Jan. 2022.

[18] Y. Yuan, J. Fang, X. Lu, and Y. Feng, "Spatial structure preserving feature pyramid network for semantic image segmentation," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 15, no. 3, pp. 1–19, 2019.

[19] Y. Yuan, H. Ning, and X. Lu, "Bio-inspired representation learning for visual attention prediction," *IEEE Trans. Cybern.*, vol. 51, no. 7, pp. 3562–3575, Jul. 2021.
[20] W. Wang and J. Shen, "Deep visual attention prediction," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2368–2378, May 2018.
[21] Z. Zheng, J. Chen, X. Zheng, and X. Lu, "Remote sensing image generation from audio," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 6, pp. 994–998, Jun. 2021.
[22] H. Ning, X. Zheng, Y. Yuan, and X. Lu, "Audio description from image by modal translation network," *Neurocomputing*, vol. 423, pp. 124–134, Jan. 2021.
[23] Q. Wang, J. Gao, W. Lin, and X. Li, "NWPU-crowd: A large-scale benchmark for crowd counting and localization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 6, pp. 2141–2149, Jun. 2021.
[24] Q. Wang, J. Gao, W. Lin, and Y. Yuan, "Pixel-wise crowd understanding via synthetic data," *Int. J. Comput. Vis.*, vol. 129, no. 1, pp. 225–245, Jan. 2021.
[25] L. Han, X. Li, and Y. Dong, "SalNet: Edge constraint based end-to-end model for salient object detection," in *Proc. Chin. Conf. Pattern Recognit. Comput. Vis. (PRCV)*. Springer, 2018, pp. 186–198.
[26] X. Li, D. Song, and Y. Dong, "Hierarchical feature fusion network for salient object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 9165–9175, 2020.
[27] H. Oliveira and P. L. Correia, "CrackIT—An image processing toolbox for crack detection and characterization," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2014, pp. 798–802.
[28] K. Gopalakrishnan, S. K. Khaitan, A. Choudhary, and A. Agrawal, "Deep convolutional neural networks with transfer learning for computer vision-based data-driven pavement distress detection," *Construct. Building Mater.*, vol. 157, pp. 322–330, Dec. 2017.
[29] 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.
[30] F.-C. Chen and M. R. Jahanshahi, "NB-CNN: Deep learning-based crack detection using convolutional neural network and Naïve Bayes data fusion," *IEEE Trans. Ind. Electron.*, vol. 65, no. 5, pp. 4392–4400, May 2018.
[31] Q. Zou, Z. Zhang, Q. Li, X. Qi, Q. Wang, and S. Wang, "Deepcrack: Learning hierarchical convolutional features for crack detection," *IEEE Trans. Image Process.*, vol. 28, no. 3, pp. 1498–1512, Mar. 2019.
[32] F. Yang, L. Zhang, S. Yu, D. V. Prokhorov, X. Mei, and H. Ling, "Feature pyramid and hierarchical boosting network for pavement crack detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 4, pp. 1525–1535, Apr. 2020.
[33] J. Liu *et al.*, "Automated pavement crack detection and segmentation based on two-step convolutional neural network," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 35, no. 11, pp. 1291–1305, 2020.
[34] B. Kim, N. Yuvaraj, K. S. Preethaa, and R. A. Pandian, "Surface crack detection using deep learning with shallow cnn architecture for enhanced computation," *Neural Comput. Appl.*, vol. 33, no. 15, pp. 1–17, 2021.
[35] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[36] M.-H. Guo, J.-X. Cai, Z.-N. Liu, T.-J. Mu, R. R. Martin, and S.-M. Hu, "PCT: Point cloud transformer," *Comput. Vis. Media*, vol. 7, no. 2, pp. 187–199, Jun. 2021.
[37] 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.
[38] 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.
[39] F. Perazzi, P. Krahenbuhl, 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.
[40] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput.-Assist. Intervent.* Springer, 2015, pp. 234–241.
[41] L. Wang *et al.*, "Learning to detect salient objects with image-level supervision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 136–145.

Jie Fang is currently with the School of Telecommunication and Information Engineering, Xi'an University of Posts and Telecommunications, Xi'an, China; and also with Corporation of Shaanxi Wukong Clouds Information and Technology, Xi'an. His research interests include imagery interpretation and multi-domain information joint perception.

Chen Yang is currently with the Ministry of Science and Technology, Pudong Development Bank, Xi'an, China. Her research interests include software engineering and information processing.

Yuetian Shi received the B.S. degree from Zhejiang University, Hangzhou, China, in 2017.

He 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, China; and also with the University of Chinese Academy of Sciences, Beijing, China. His research interests include machine learning, image processing, and point cloud.

Nan Wang received the B.S. degree from Zhejiang University, Hangzhou, China, in 2018.

He 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, China; and also with the University of Chinese Academy of Sciences, Beijing, China. His research interests include computer vision and deep learning.

Yang Zhao received the B.E. and M.E. degrees in transportation engineering from Chang'an University, Xi'an, Shaanxi, China, in 2014 and 2017, respectively, and the Ph.D. degree from the School of Transportation and Logistics, Southwest Jiaotong University, Chengdu, Sichuan, China, in 2021.

He is currently a Lecturer with the College of Transportation Engineering, Chang'an University. His research interests include metro operation and management, machine learning, and data mining.