Self-adaptive 2D-3D Image Fusion for Automated Pixel-level Pavement Crack Detection

Jiayv Jing [a,b] , Xu Yang [a,b,*] , Ling Ding [c] , Hainian Wang [a], Jinchao Guan [d], Yue Hou [e] , Sherif M. El-Badawy [f]

[a] School of Highway, Chang'an University, Xi'an, 710064, China

[b] College of Future Transportation, Chang'an University, Xi'an, 710064, China

[c] Collage of Transportation Engineering, Chang'an University, Xi'an 710064, China

[d] Zhejiang Expressway Co., Ltd., Hangzhou 310020, Zhejiang, China

[e] Department of Civil Engineering, Faculty of Science and Engineering, Swansea University, UK.

[f] Public Works Engineering Department, Faculty of Engineering, Mansoura University, Mansoura 35516, Egypt

ABSTRACT

Current 2D and 3D image-based crack detection methods in transportation infrastructure often struggle with noise robustness and feature diversity. To overcome these challenges, the paper use CSF-CrackNet, a self-adaptive 2D-3D image fusion model utilizes channel and spatial modules for automated pavement crack segmentation. CSF-CrackNet consists of four parts: feature enhanced and field sensing (FEFS) module, channel module, spatial module, and semantic segmentation module. A multi-feature image dataset was established using a vehicle-mounted 3D imaging system, including color images, depth images, and color-depth overlapped images. Results show that the mean intersection over union (mIOU) of most models under the CSF-CrackNet framework can be increased to above 80%. Compared with original RGB and depth images, the average mIOU increases with image fusion by 10% and 5%, respectively. The ablation experiment and weight significance analysis further demonstrate that CSF-CrackNet can significantly improve semantic segmentation performance by balancing information between 2D and 3D images.

Keywords： Pavement crack detection; Self-adaptive image fusion; Semantic segmentation; Multi-feature dataset

1. Introduction

Cracks may cause significant failure to road surface infrastructures. In the past, the manual vision detection method was widely used for road distress detection that was unable to meet the mass inspection amount of the current in-service highways. Thus, researchers are paying more and more attention to the automatic detection of pavement distresses. While deep learning has undoubtedly made significant contribution to the field of image segmentation, its application in the real-world road engineering projects, faces considerable challenges. The diverse range of road distresses and the complexity of road environment are posed significant obstacles to accurately segmenting road cracks using advanced methods.

2D RGB images have been used by many existing pixel-level crack recognition researches using deep learning because 2D RGB images are easy to obtain. However, in real-world road engineering cases, the shadows on the road, water stains, and wheel path on the road will cause the crack pixel level segmentation task to become particularly difficult to execute accurately. One of the possible problem-solving method, deep learning approach can be used in the road distress identification process [1]. Xu et al.[2] proposed a two-stage pavement distress image enhancement pattern for dataset expansion to improve the richness of data. The distress prediction performance was improved by increasing the number of complex samples. Ren et al. [3]proposed a semi-supervised learning approach based on generative adversarial networks for identifying pixel-level anomalous image segments. This method can reduce the workload of data annotation, thus providing a richer data form for deep learning networks. Furthermore, researchers have extensively investigated 2D data analysis pertaining to pavement distress detection. Zhang et al. [4] proposed a framework for asphalt pavement distress detection called ShuttleNetV2, with capabilities of enhanced global modeling and retrieval of fine details. Tong et al.[5] proposed a deep neural network combining the Dempster-Shafer theory (DST) and a transformer network. The excellent information extraction ability of the transformer also improves the pixel-level detection accuracy of the road surface. Lin et al. [6] proposed GoogleNet transfer learning with an improved gorilla optimized kernel extreme learning machine. Through transfer learning and graphics preprocessing, the poor detection accuracy of high noise images has been effectively improved. Optimized information extraction methods and image preprocessing can also improve the detection accuracy of simple 2D image targets. However, current research still struggles to address the challenges posed by 2D images greatly affected by sever environmental conditions and poor illumination. Most existing research assumes of ideal conditions, overlooking the complexities exist in the real-world.

Compared with 2D images, 3D images are less easily affected by environmental conditions. It provides more effective information for distress segmentation while reducing image noise. In 2017, Zhang et al. [7] developed a convolutional neural network architecture, CrackNet, for pixel-level crack detection in 3D pavement images. In order to eliminate the influence of local noise on crack prediction results, Zhang et al. [8] proposed CrackNet II using a deeper network structure in 2018. Both of them proved that 3D images can perform well in pavement crack detection tasks and can effectively reduce the interference of environmental factors on pavement crack prediction. However, various pavement forms and the complexity characteristics of pavement distress are still the reasons for the low accuracy of pre-distress prediction. Fei et al. [9] proposed an improved CrackNet called CrackNet-V for pixel-level automated crack detection on 3D asphalt pavements in 2020. Even after many improvements, CrackNet-V still faces the problem of inaccurate detection of wide cracks. This is because the wider cracks will be filled with fine impurities such as sand, which can easily cause the cracks to be discontinuous in 3D space. Liu et al. [10] proposed a hybrid method to automatically detect inverted-T patching for an efficient maintenance schedule. However, they found that the inverted-T patching and background in 3D image are so similar, which is the main cause of false-positive.

Therefore, 3D images frequently encounter issues in areas where there is minimal variation in the height. For instance, during the early stages of crack development (micro-cracks), as well as in the cases of co-developed distresses (cracks in local subsidence areas), and repaired distresses (cracks after filling), automatic detection often yields poor results.

Deep learning has proved to be effective in image processing, however, the inherent limitations of using 2D or 3D images alone persist and cannot be fully addressed. Compared with other engineering cases, road images are much more difficult to process, which requires complicates subsequent data processing, distress analysis, information extraction and more. In practical applications, pavement crack images have the characteristics of irregularity, diversity of structural surface, variability of environment and uncertainty caused by non-crack features. In order to enrich the dimension of image information and increase recognition accuracy, the image fusion algorithm is often employed. In recent years, deep learning methods have shown great potential in the field of image fusion[11], among which convolutional neural networks (CNNs) have gradually become the main tool for image fusion. Prabhakar et al. [12]used a convolutional neural network to extract the information of the image in the brightness channel. Based on ResNet50, Li et al. [13]fully extracted the features of the source image to realize the fusion algorithm of the infrared and visible images. However, different data types have also different fusion strategies. In the field of pavement distress detection, Guan et al.[14] established a multi-feature pavement image dataset including color image, depth image and fusion image, and discussed the possibility of fusing 2D and 3D images to improve segmentation performance. Bavirisetti et al. [15]devised an adaptive thresholding technique that utilizes local image statistics for improved segmentation of MRI scans, thereby facilitating more accurate medical diagnoses. Heideklang et al. [16] integrated three different data types through heterogeneous data fusion to improve detection performance. Beckman et al. [17] developed a concrete spalling damage detection method based on convolutional neural network using 2D and 3D images data. Zhang et al. [18] introduced a method leveraging a wavelet-based fusion technique to integrate global and local image features, enhancing underwater images with remarkable fidelity. Mouaddib et al. [19] employed a dual-method approach to assess the structural integrity of Notre-Dame's vaults by integrating 2D photogrammetric data and 3D laser scanning data, demonstrating the necessity of multi-temporal data fusion for precise structural diagnosis. At present, most of the fusion strategies are based on simple fixed formulas, resulting in poor fusion performance. In the field of pavement distress segmentation, there is also a lack of a fusion scheme combining the characteristics of cracks. Li et al. [20] proposed a method for detecting self-fusion pavement images based on convolutional neural networks. Jones et al. [21] introduced an innovative technique for enhancing the resolution of satellite images by employing a deep learning-based super-resolution framework. Zhao et al. [22] introduced a novel coarse-to-fine LiDAR and camera fusion-based network, named LIF-Seg, to address the challenges of effective fusion and precise alignment of LiDAR and camera data for 3D semantic segmentation. Considering the significant progress, the fusion process of this model still depends on the ability of the neural network to extract information. This will make it difficult to deploy the network on the mobile devices in a lightweight design. It is also difficult to migrate the method to the real-world engineering applications. Because of the lack of fusion network optimization for the essential characteristics of pavement distresses, the method performs poorly for pavement cracks with high requirements for edge information extraction. However, it has been fully demonstrated that multi-dimensional information fusion can significantly increase information density and improve detection levels.

In general, from the perspective of enriching image data, fusing multi-source images are more effectively than using homogeneous data from a single source. Since the 2D RGB image can provide rich real-world color information and reflect the plane gap between the pavement crack and the background, especially in the local high depth change area. The 3D depth image can ignore the road noise caused by poor illumination conditions, thus it can more accurately reflect the road texture and crack shape information. It can also significantly improve the accuracy of crack segmentation. Thus, in this research, a multi-dimensional dataset of road cracks is constructed, and each image contains four RGB channels and depth. An adaptive 2D and 3D image fusion called CSF-CrackNet is proposed, which can be flexibly deployed at the front end of any semantic segmentation network to improve the network detection accuracy significantly. CSF-CrackNet aims to improve the accuracy and robustness of pavement crack segmentation by utilizing a self-adaptive 2D-3D image fusion mechanism. This approach integrates the rich color information from RGB images with the structural details from depth images, dynamically adjusting weights for different image channels and spatial regions. This fusion effectively mitigates issues such as shadows, varying lighting, and fine detail loss, enhancing segmentation precision across diverse real-world scenarios. The model employs several innovative modules to enhance feature extraction, spatial weighting, and channel fusion, ensuring superior performance under challenging conditions. CSF-CrackNet is designed for flexible integration with various semantic segmentation networks, demonstrating significant performance improvements and broad applicability in real-world pavement crack detection tasks. The paper scrutinized the effects of varying input data on the model, and a comparative analysis of the proposed methods was carried out.

2.Methodology

CSF-CrackNet is a deep learning model with an encoder-only architecture optimized for pavement crack segmentation through a self-adaptive 2D-3D image fusion mechanism. The model integrates RGB and depth image data using specialized channel and spatial information analysis modules, which dynamically adjust to optimize feature capture and integration from both image types. These modules employ advanced convolution techniques, such as dilated and transposed convolutions, to enhance the processing of multiscale features critical for accurate segmentation. The fusion of channel and spatial data is designed to maximize the complementary attributes of RGB and depth information, improving the model's accuracy and robustness across diverse environmental conditions.

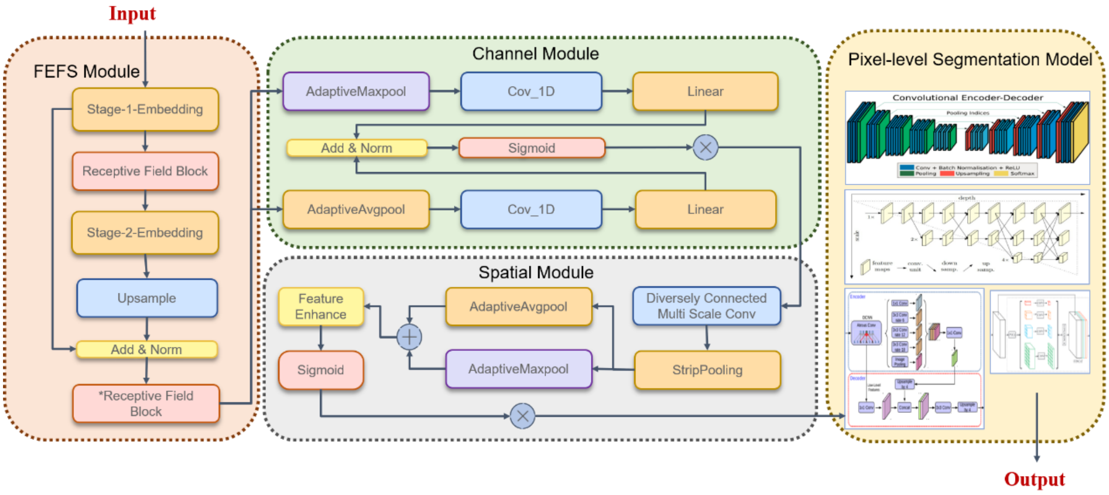


Figure 1. Overview of CSF-CrackNet(CSF-X) model

Figure 1 illustrates the innovative architecture of CSF-CrackNet, highlighting the adaptive channel and spatial fusion modules. These modules are crucial for dynamically integrating RGB and depth information, setting the model apart from traditional fixed fusion approaches. To achieve better adaptivity in pixel-level crack detection tasks, the encoder-only architecture in this paper can be divided into four parts: feature enhanced and field sensing model (FEFS), channel module, spatial module, and semantic segmentation model. Firstly, the receptive field block (RFB) and the shortcut pattern are combined to extract whole deep crack information, expand the receptive field, and summarize latent representations. Secondly, the channel feature maps from RGB images and depth images are reasonably applied to maps of different weights, and the intermediate features are adaptively refined. Thirdly, the information extraction module is added to the space module again and then recombines and strengthens the spatial features because the channel model re-processes the feature maps. Finally, the images after channel and spatial fusion are input into the semantic segmentation network for crack segmentation. In addition, there is no limitation of the semantic segmentation network used in this framework, indicating that the fusion model can be easily deployed before any semantic segmentation network. Overall, the purpose of this framework is to improve segmentation accuracy through the fusion of multimodal features, where the architecture and function of each module are described in the following sections.

2.1 Feature enhancement and field sensing module

The FEFS module can generate feature maps with richer information by multilayer convolutional network operation. The depth is beneficial for the accuracy of information processing [23]. Therefore, the primary function of FEFS is to obtain a deeper feature map. By expanding the receptive field, rich contextual information can be effectively obtained. Additionally, using a larger convolution kernel or a larger pooling step size can increase the receptive field of the network [24]. As shown in Figure 2, the proposed improved Receptive Field Block (RFB) [25] in this study not only integrates the inception structure with dilated convolution layers but also introduces a novel multi-branch configuration tailored specifically for crack detection. This configuration enhances the capture of fine-grained details and long-range dependencies, crucial for detecting narrow and continuous road cracks. Additionally, by incorporating adaptive skip connections, our RFB mitigates the potential over-amplification and weakening of responses, thereby maintaining stable and enhanced low-level feature representations. This refinement over traditional RFB designs makes our approach uniquely suited for the complexities of pavement crack detection. In addition, the jump connection can avoid over-amplification and over-weakening of the response between any two channels. And it can retain the representation level of low-level features [26]. Therefore, the model structure with skip connection is utilized for stable feature enhancement.

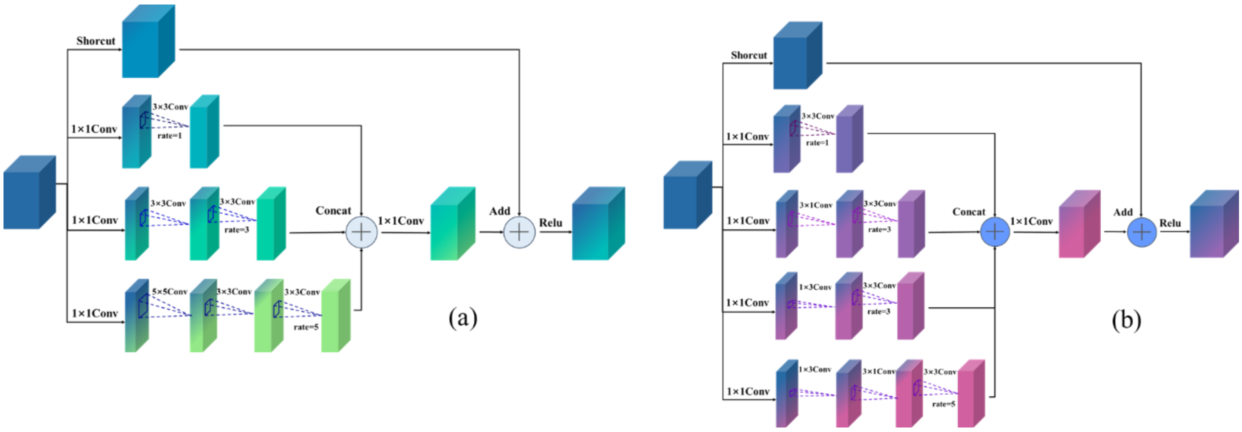


Figure 2. Architecture of the receptive field block (RFB) module. (a) RFB model (b) RFB-small model

As shown in Figure 1, the image is first input into the embedded block with double convolutional layers, as shown in Figure 3. The 3 × 3 max pooling is also performed for downsampling until a quarter of the original size of the multi-channel simple semantic information is obtained [27]. Subsequently, the obtained feature map is copied and input into two branches. One branch maintains its course through a direct jump connection, preserving the fundamental simple information. Another branch is transferred into RFB. The feature map emerging from the RFB block is then reintroduced into the embedding block, generating deeper semantic information. This process is repeated until multi-channel simple semantic information, reduced to one-eighth of the original size, is achieved. Next, the deep semantic information is inputted into the upsampling layer to enlarge feature maps to a quarter of the original to display them at higher resolution. Afterwards, these feature maps are concatenated with the previously branch-retained feature maps in the skip connection and then regularized. Ultimately, these feature maps are inputted into RFB-s to obtain enhanced feature information with multi-dimensional information.

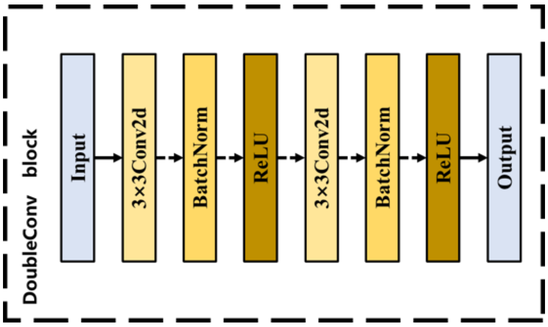


Figure 3. Double-layer convolution architecture

Figure 2 describes two RFB configurations involving a multi-branch convolution layer in tandem with either dilated pooling or convolution layers [25]. The initial step involves the reduction of channel count in the input feature maps through a 1×1 convolution to facilitate information aggregation. Subsequently, a series of convolution and dilation convolution operations transpire across multiple branches. Thereafter, the feature maps generated from these branches are concatenated along the channel dimension, followed by a 1×1 convolution to restore the original channel feature map. The resulting output is augmented with the shortcut outputs. This summation undergoes nonlinear activation through the Rectified Linear Unit (ReLU) to produce the final output. The aforementioned steps delineate the comprehensive process of the Receptive Field Block (RFB). Notably, the framework introduces RFB-s (depicted in Figure 2(b)), incorporating smaller convolution kernels and additional branches in the network to meticulously analyze the characteristics of fine and small cracks.

2.2 Channel feature fusion based on one-dimensional convolution

The channel feature fusion module applies one-dimensional convolution to automatically learn and adjust the significance of each channel in RGB and depth images. This adaptive weighting mechanism enhances the emphasis on critical features while minimizing less relevant information, refining the model's focus and improving segmentation performance. Unlike traditional channel fusion mechanisms that rely on fixed formulas or predefined rules, our method introduces a novel approach to channel weight adaptation through one-dimensional convolution and fully connected layers. This approach allows the network to dynamically learn and adjust the importance of each channel based on the specific characteristics of the input images. By autonomously acquiring feature weights during training, our method ensures optimal feature extraction tailored to each image, enhancing robustness and accuracy in crack detection tasks. This process amplifies the weight assigned to more impactful feature channels, enhancing the network's ability to prioritize and leverage effective features. In this way, each sample will have its own independent set of weights. For instance, the weights of any two image samples can be adjusted adaptively according to the image quality.

Pooling is a common operation in convolutional neural networks, also known as downsampling, which aims to reduce the dimension of each feature map[28]. Therefore, as shown in Figure 1, at the beginning of this module, the feature maps processed by the FEFS module are divided into two branches and input into the Max pooling layer and Average pooling layer, respectively, with one-dimensional output. After processing, the output of the feature maps of two branches is a multi-channel one-dimensional graph vector.

Following this, the feature maps are fed into a data analysis block that incorporates one-dimensional convolution and fully connected layers. This block extracts information and condenses features from the input feature map. As depicted in Figure 4, a crucial step involves smoothing and denoising the data processed by the pooling layers. The data values undergo compression, resulting in a 1024-dimensional vector that is subsequently normalized. Utilizing standardized image data, a one-dimensional convolution layer with a kernel length of 7 is employed to extract local features from the preprocessed vector. Generating 40 feature vectors, each with a length of 1024 dimensions. Subsequently, these local features undergo abstraction through another one-dimensional convolution layer with a kernel size of 5, effectively reducing the number of feature vectors to one quarter. After three layers of one-dimensional convolution operations, four feature vectors, each with a length of 1024 dimensions, are extracted and then flattened to generate a comprehensive feature vector of 4096 dimensions, preparing it for full connection layer processing. The subsequent step involves concatenating two branches that employ different pooling methods. Feature compression occurs through five fully connected layers, incorporating the Rectified Linear Unit (ReLU) activation function. Following this, the Sigmoid function is applied for activation, resulting in the output of the channel weight mask. Finally, the weight and input 2D-3D images undergo channel-wise multiplication, yielding the ultimate refined feature maps.

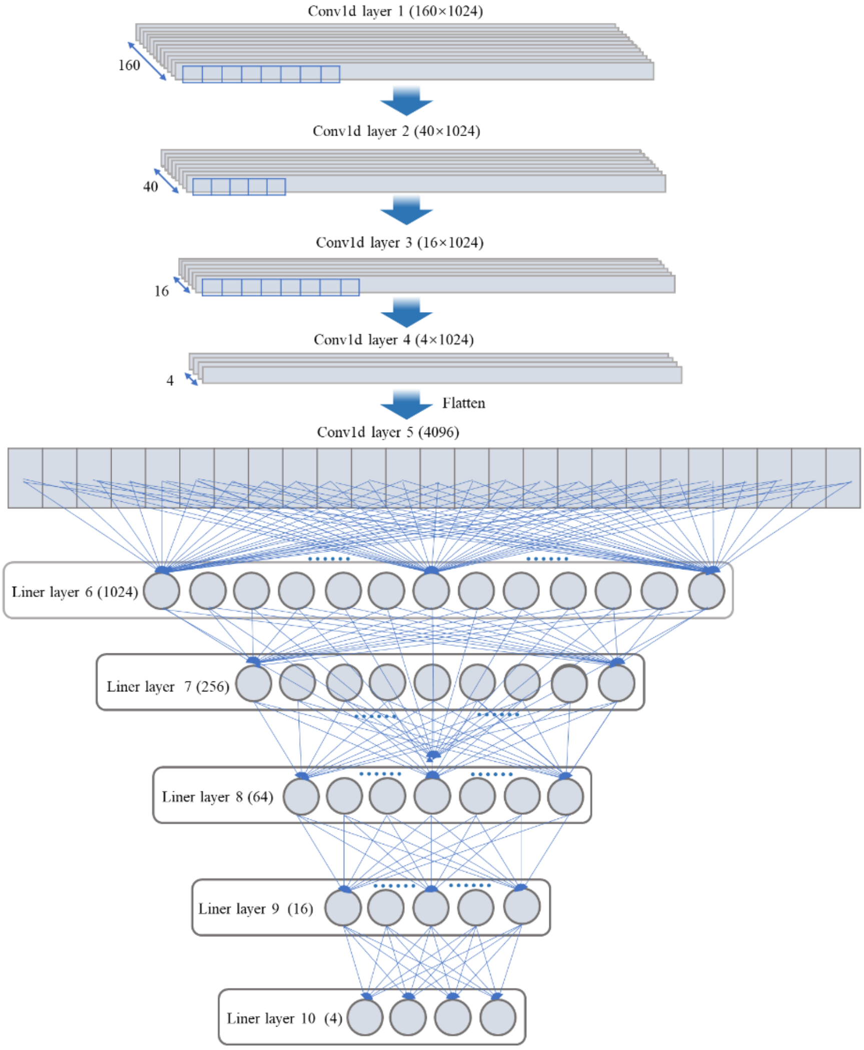


Figure 4. Architecture of one-dimensional convolution and linear fully connected layer

2.3 Spatial feature fusion module with multi-scale features and Scene Parsing

The Spatial Feature Fusion Module selectively enhances features in key areas for crack segmentation by transforming and refining spatial information. This module generates a spatial weight mask for each position, adjusting the emphasis on relevant regions and diminishing background noise. Initially working with shallow features from the channel fusion, it abstracts these to deeper semantic levels for more precise segmentation. To achieve this, the model incorporates diversely connected multi-scale convolution blocks and Strip pooling blocks, which process images post-channel fusion to enhance detail representation and scene parsing.

Although the low-level semantic feature information is less, the target location is clear. The high-level semantic feature information has opposite characteristics. The spatial pyramid structure fuses the features of different layers with low-level and high-level semantic information to achieve better results[29]. Therefore, as Figure 5 shows the proposed structure of a diversely connected multi-scale convolution block, this block uses the feature pyramid structure to introduce region of interest pooling and transposed convolution for feature map abstraction. The block comprises two consecutive down-sampling operations utilizing ROI pooling, followed by two additional down-sampling steps facilitated by transposed convolution. By integrating region of interest pooling with transposed convolution, we achieve a more granular abstraction of feature maps, allowing for precise detection of various crack scales and forms. The block employs a dynamic feature pyramid structure that adaptively adjusts to different crack widths and patterns, ensuring robust performance across diverse pavement conditions. This design not only improves detection accuracy but also enhances computational efficiency, making it highly effective for real-time applications in road maintenance and monitoring.

The fusion of semantic information across various depths is accomplished by concatenating feature maps from different levels.

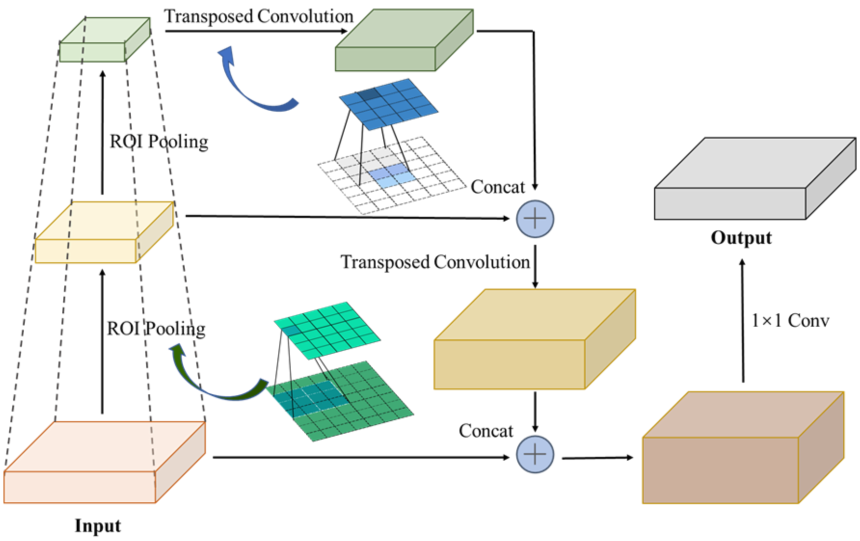


Figure 5. Architecture of Diversely connected multi-scale convolution block

Spatial pooling has been proven highly effective in capturing long-range contextual information for pixel-wise prediction tasks[30]. In Figure 6, different from the traditional N × N style convolution kernel, strip pooling introduces a novel pooling strategy that involves a long yet narrow kernel, specifically 1 × N or N × 1. This elongated pooling window allows the model to gather abundant global contextual information, a crucial aspect for enhancing the performance of scene parsing networks.

Additionally, by incorporating dilated convolutions within the strip pooling framework, we significantly expand the receptive field, allowing the model to integrate more comprehensive scene context without increasing computational burden. This dual enhancement of spatial pooling and depthwise separability sets our method apart from conventional strip pooling techniques, delivering superior performance in pixel-level segmentation of complex crack patterns.

The operation of spatial fusion is similar to channel fusion. The abstract deep semantic information is input into the maximum pooling layer and the average pooling layer, respectively, to obtain each spatial position's maximum and average values. After obtaining two matrices, the two matrices are concatenated. The model learns the weight mask for each spatial position by applying a convolutional layer and the sigmoid function. In the final step, this weight mask is applied to each feature map's spatial position, emphasizing and highlighting crucial information.

Our spatial feature fusion module introduces an innovative self-adaptive spatial weighting mechanism that leverages multi-source information from both RGB and depth images. By combining diversely connected multi-scale convolution blocks and enhanced strip pooling, our approach dynamically adjusts spatial weights to emphasize crack regions and suppress irrelevant background noise. This fusion of multi-source information ensures that the most critical features from both RGB and depth images are prioritized, significantly improving the precision of crack segmentation. The RGB images provide rich color and texture data, which is essential for identifying surface characteristics and crack edges under varying lighting conditions. However, they can be affected by shadows and other environmental factors. On the other hand, depth images offer structural details and depth information that are less susceptible to lighting variations, providing a complementary perspective that enhances the overall robustness of the segmentation process. By integrating these two types of information, our module captures a wider range of contextual data, crucial for accurate crack detection. Traditional single-source methods struggle to achieve the same level of detail and robustness, as they cannot simultaneously address the challenges posed by varying lighting conditions and the need for structural depth information. Our multi-source approach ensures that the segmentation model benefits from the strengths of both image types, resulting in a more comprehensive and reliable detection framework.

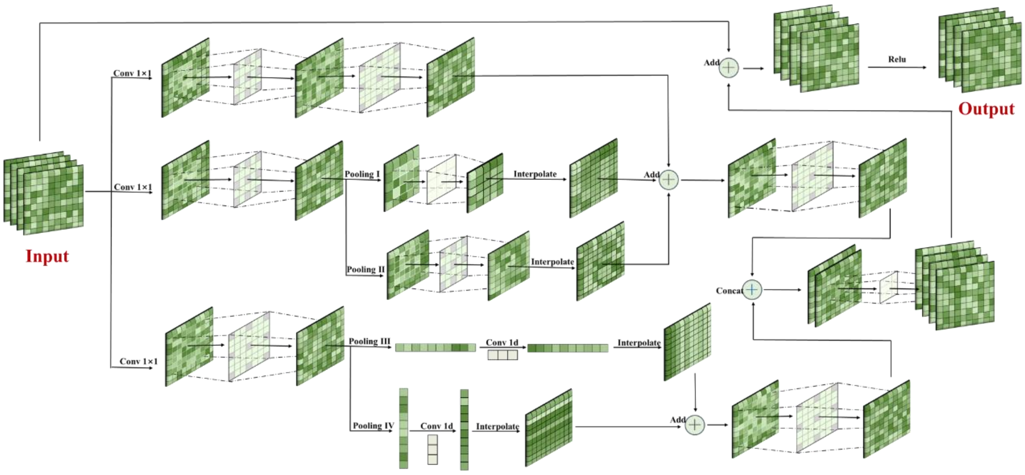


Figure 6. Architecture of improved Strip Pooling based on Depthwise Separable Convolution

2.4 Semantic segmentation model

Following spatial fusion, the resulting 2D-3D fusion image undergoes semantic segmentation for detailed crack analysis. The model is designed to seamlessly integrate with existing semantic segmentation architectures, including Deeplab V3+, Unet, PSPnet, HRNet, and Segnet, enhancing their performance without the need for structural modifications. This compatibility ensures that the sophisticated feature processing capabilities of CSF-CrackNet can be utilized across various platforms to achieve precise pixel-level crack segmentation.

2.5 Previous fusion methods used for comparison

In the realm of image fusion, Multi-scale Guided Filter Fusion (MGFF) and Convolutional Neural Network (CNN)-based methods stand out due to their widespread application and exceptional capabilities in enhancing image quality. Therefore, this paper selects these two methods as comparative algorithms.

2.5.1 Multi-scale Guided Filter Fusion

The multi-scale guided filter fusion (MGFF) integrates information from different source images using a guided image filter (GF) and advanced techniques such as multi-scale image decomposition, visual saliency detection, and structure transferring property[15]. By combining pixel-level details from various sources, the algorithm ensures a comprehensive representation in the fused image or video. Through multi-scale decomposition, the algorithm extracts feature at different levels of detail, preserving important information during fusion. Visual saliency detection identifies significant regions in the source images, focusing on key areas for preservation. The structure transferring property transfers structural information from source images to maintain coherence in the final output. Weight maps guide the fusion process based on the importance of different regions. Overall, the algorithm aims to maximize fusion gain, minimize loss and artifacts, and optimize run time. This results in efficient and high-quality fused images and videos for applications in diverse fields like robotics, surveillance, and medical imaging[31-33].

2.5.2 Fusion based on convolutional neural networks

The fusion using convolutional neural networks (CNN) involves a multi-step process[21]. Firstly, a Siamese convolutional network generates a weight map by processing the images separately. This weight map integrates pixel activity information from both images. To handle images of arbitrary sizes, the fully-connected layer of the network is converted into an equivalent convolutional layer with two kernels. This allows the network to process source images as a whole and generate a dense prediction map containing clarity information for each patch pair. The network output simplifies to the weight of the first or second source image. Finally, a weight map with the same size as the source images is obtained by assigning weights to all pixels within the patch locations and averaging the overlapped pixels. This fusion scheme ensures that the fusion process is conducted multi-scale, adapting the fusion mode for decomposed coefficients based on local similarity, ultimately achieving high-quality fusion results. This method is widely used in agriculture, computer vision and other fields with excellent image fusion performance [34-36].

2.5.3 Comparison between the proposed method and the previous methods

The mentioned image fusion methods primarily include approaches based on multi-scale decomposition and sparse representation. Different fusion methods rely on the selection of image decomposition techniques and the formulation of fusion rules. The core of image fusion lies in obtaining weight maps that capture significant information from each source image. This crucial step is achieved through saliency level estimation and weight allocation. Methods based on convolutional neural networks are constrained by network structures and lack optimization for image fusion algorithms in complex road scenarios. While conventional fusion methods have shown promising application results, several technical challenges urgently need to be addressed. Firstly, the limitation lies in manually designed fusion rules, leading to insufficient robustness in image fusion effects. Secondly, efficiency is compromised in the case of complex and diverse datasets. Thirdly, for road crack problems, there is a lack of fusion strategies specific to road defect features and a shortage of fusion algorithms tailored to road scenes.

CSF-Cracknet is an adaptive graphic fusion algorithm built upon a finely multi-source dataset of road cracks. Addressing the characteristics of both fine and large-scale cracks on road surfaces, it introduces a pyramid-structured surface feature map abstraction unit. A scene-awareness module is proposed to account for the diversity in road surface textures. A weight-based fusion strategy is presented in response to road surface occlusion and shadow issues. The aim is to achieve a high-confidence extraction of road texture features and robust pixel-level segmentation of road cracks.

3. Data preparation

3.1 Data collection and processing

This study used a 3D imaging system developed by our research team[14] . The vehicle-mounted photography system based on multi view stereo imaging technology was used to generate the digital pavement surface model. Based on a high-resolution point cloud model, a multi feature image dataset consisting of color images, depth images, and color-depth overlapped images was created using image processing algorithms.

The dataset collection utilizes a vehicle-mounted photography system with several GoPro cameras to capture pavement images. Camera calibration is performed to eliminate lens distortion. The images are processed using structure from motion (SfM) technology to reconstruct a 3D point cloud model. The point cloud model is transformed into orthoimages by a Python script with batch image processing. This comprehensive approach ensures acquiring and processing a high-quality dataset for automated pixel-level pavement distress detection. For the 2D color images, the RGB values of each pixel on the image are represented. In contrast, each pixel on the 3D depth image represents the average height of the point cloud within the region. Both types of orthophoto images utilize the same data source and share an identical imaging range. Therefore, the generated two-dimensional and 3D images exhibit complete overlap characteristics.

3.2 Pavement crack multi-dimensional dataset

The 900 sets of pavement crack multi-dimension datasets were used in this paper by the 3D imaging system. The 900 datasets were randomly divided into 700 training sets, 100 validation sets, and 100 testing sets. Each dataset consists of three images: an RGB image, a depth image, and a Ground Truth image. Crack distresses primarily manifest in linear and grid-like forms. Among them, the longitudinal and transverse cracks exhibit relatively regular patterns, while block-like and grid-like cracks typically intertwine with multiple cracks in images. These sets were utilized as the source data to train and assess various deep learning networks.

The 3D image fundamentally differs from the 2D image in expressing detailed road surface information by representing distance and depth, offering a comprehensive depiction of crack location, depth, and shape. In contrast, the 2D image conveys color information characterizing surface brightness and providing details on color and texture. Both the road depth information and road color information can be expressed in a 2D matrix. However, the 2D image of the road surface differs from the 3D image, but they are interrelated. Because of the great complementarity between the two types of images, the efficient fusion of the two images can make up for the defects between the two and make it more accurate for feature extraction and recognition.

To ensure that the damage identification method can adapt to real road surface scenarios, the road surface damage image dataset incorporates various complex road environment conditions. Road surface color, lighting shadows, and surface stains significantly impact the robustness of damage identification. The usage conditions of roads affect the contrast and color difference between damaged and non-damaged areas. The road surface damage images include mildly worn surfaces (tending to black), heavily worn surfaces (tending to gray), and surfaces with surface floating dust (tending to yellow). On the other hand, considering the intensity and angle of illumination can affect the visual conditions of the road surface, and shadows cast by trees or buildings can lead to irregular color difference distributions. Additionally, surface stains such as oil stains, water stains, and repairs are complex interference factors. Under various combinations of external interferences, the road surface damage image dataset comprehensively tests the stability of subsequent recognition algorithms in various real-world scenarios. Figure 7 illustrates several representative matched sets of 2D and 3D images, including various complex noises such as shadow, water stains, road marking, wheel paths, and local subsidence.

In addition, the size of the pavement crack multi-dimension images is 512 × 512 (H × W) pixels. Each dataset includes a 3D pavement image, a paired 2D image and a ground-truth image aligned on a pixel-to-pixel basis. All ground-truth images underwent manual labeling using the LabelMe[37].

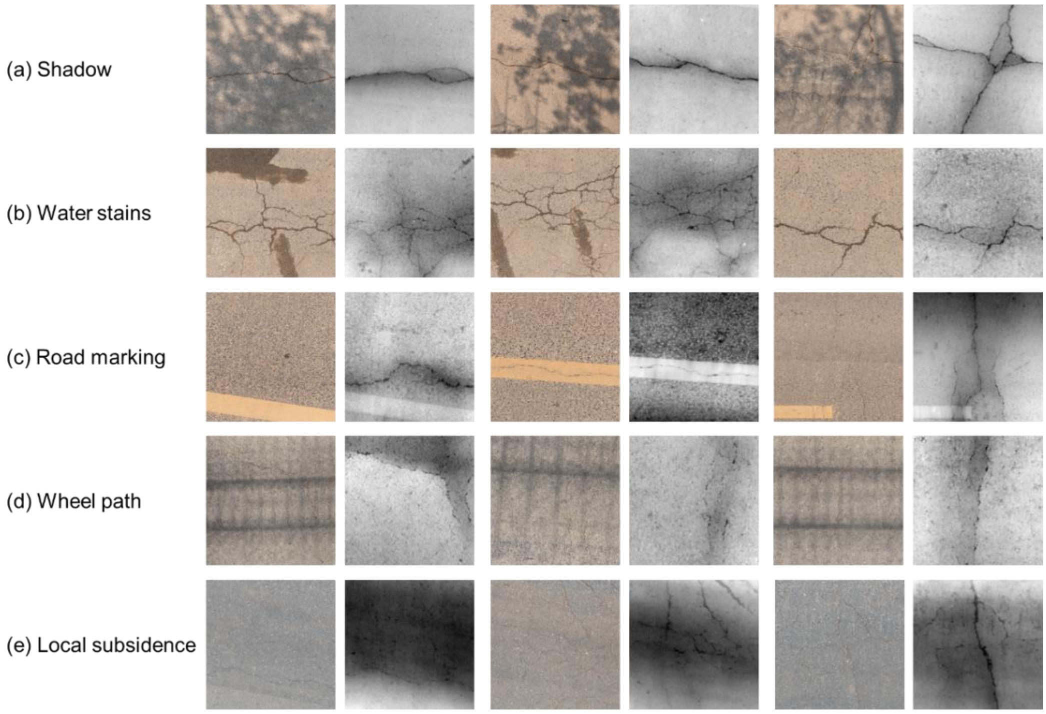


Figure 7. Partially representative 2D-3D images-

4. Experimental results and performance comparison

4.1 Evaluation of segmentation performance

4.1.1 Benchmarking experiments and models

In this paper, two sets of comparative experiments are carried out using CSF-CrackNet to ascertain its superiority over other fusion and non-fusion methods, along with its compatibility with different segmentation networks. To prove that the images processed by CSF-CrackNet are more conducive to crack segmentation, the paper uses different kinds of data to deploy in the same model framework for comparative experiments, including the fusion images based on convolutional neural network and feature pyramid (CNN)[38], the fusion images based on Multi-scale Guided Filter Fusion (MGFF)[39], RGB images, depth images and 2D-3D images. To demonstrate that the CSF-CrackNet model can be flexibly deployed at the front end of any semantic segmentation network to improve the network detection accuracy significantly, the paper attempts to deploy CSF-CrackNet to the front end of multiple mainstream semantic segmentation models for testing, including DeepLab V3+[40], Unet[27], PSPNet[41], HRNet[42] and SegNet[43]. For a fair comparison, all these networks are trained with the same hyperparameters mentioned above. In the following subsections, the evaluation results of CSF-CrackNet are described in detail.

4.1.2 Quantitative comparison of different models

Table 1 describes our experiments to verify the good performance of the model in CSF-CrackNet. We also performed similar experiments based on DeepLab V3 +, PSPNet, and HRNet, for a total of 25 sets of experiments that combine various models and data for comparative analysis.

Table 1 Description of the models to be trained

Framework	Model name	Description	Dataset of training
	Deeplab V3+ 2D	Original Deeplab V3+ network	RGB images

Deeplab V3+	Deeplab V3+ 3D	Original Deeplab V3+ network	Depth images
	MGFF-Deeplab V3+ 2D+3D	Fused images based on MGFF fusion is segmented by Deeplab V3+ network.	RGB & Depth images
	CNN-Deeplab V3+ 2D+3D	Fused images based on CNN is segmented by Deeplab V3+ network.	RGB & Depth images
	CSF-Deeplab V3+ 2D+3D	Method based on channel and space fusion proposed in this paper is deployed in the front of Deeplab V3+ network.	RGB & Depth images
Unet	Unet 2D	Original Unet network	RGB images
	Unet 3D	Original Unet network	Depth images
	MGFF-Unet 2D+3D	Fused images based on MGFF fusion is segmented by Unet network.	RGB & Depth images
	CNN-Unet 2D+3D	Fused images based on CNN is segmented by Unet network.	RGB & Depth images
	CSF-Unet 2D+3D	Method based on channel and space fusion proposed in this paper is deployed in the front of Unet network.	RGB & Depth images
PSPnet	PSPnet 2D	Original PSPnet network	RGB images
	PSPnet 3D	Original PSPnet network	Depth images
	MGFF-PSPnet 2D+3D	Fused images based on MGFF fusion is segmented by PSPnet network.	RGB & Depth images
	CNN-PSPnet 2D+3D	Fused images based on CNN is segmented by PSPnet network.	RGB & Depth images
	CSF-PSPnet 2D+3D	Method based on channel and space fusion proposed in this paper is deployed in the front of PSPnet network.	RGB & Depth images
Hrnet	Hrnet 2D	Original Hrnet network	RGB images
	Hrnet 3D	Original Hrnet network	Depth images
	MGFF-Hrnet 2D+3D	Fused images based on MGFF fusion is segmented by Hrnet network.	RGB & Depth images
	CNN-Hrnet 2D+3D	Fused images based on CNN is segmented by Hrnet network.	RGB & Depth images
	CSF-Hrnet 2D+3D	Method based on channel and space fusion proposed in this paper is deployed in the front of Hrnet network.	RGB & Depth images
Segnet	Segnet 2D	Original Segnet network	RGB images
	Segnet 3D	Original Segnet network	Depth images
	MGFF-Segnet 2D+3D	Fused images based on MGFF fusion is segmented by Segnet network.	RGB & Depth images
	CNN-Segnet 2D+3D	Fused images based on CNN is segmented by Segnet network.	RGB & Depth images

CSF-Segnet 2D+3D	Method based on channel and space fusion proposed in this paper is deployed in the front of Segnet network.	RGB & Depth images

Figure 8 illustrates the loss and mIOU of different models on the validation images during the training process. Figure 8 (a) shows the loss curve of different datasets mentioned in section 4.2.1 using the Unet model on the validation dataset. It can be seen from the figure that as the number of iterations increases, the loss gradually decreases, indicating that the performance of the model is gradually improving. Simultaneously, the number of iterations required for different network structures to achieve the same performance is also different. The model (CSF-Unet 2D+3D) proposed in this paper requires fewer iterations to achieve lower losses. This means that these network structures perform better when dealing with crack segmentation. Figure 8 (b) shows the loss curves of different datasets using the SegNet model on the validation dataset. Like the results of the Uent model, the loss curve of the model (CSF-Segnet 2D+3D) proposed in this paper decreases the fastest. However, the loss value of the fused image dataset using MGFF is larger than that of the Unet model. This means that the fusion method of MGFF shows unstable performance when fusing crack RGB images and depth images. The model (CSF-Segnet 2D+3D) proposed in this paper is more robust and has a faster convergence speed in the training process. This is because good image fusion results can help the network capture the deep information of the graph faster and more accurately.

Figure 8 (c) and 8 (d) describe the mIOU performance results of different datasets and model, which show similar characteristics. It is evident from the figure that the model (CSF-Segnet 2D + 3D) proposed in this paper shows the best results. The mean intersection over union ratio at stability exceeds 0.8, and the convergence speed is also the fastest. Our proposed model mIOU exceeds the training results (Segnet 2D) of the RGB image dataset by about 8 % and exceeds the training results (Segnet 3D) of the depth image dataset by about 3 %. This shows that the fusion method realizes the extraction and enhancement of the effective information of the image, which is helpful for the segmentation of pavement cracks. Furthermore, the fusion performance is better than other fusion methods. It is worth mentioning that the MGFF fusion method (MGFF-Segnet 2D + 3D) has a harmful effect on the segmentation performance results, which further illustrates the importance of combining the fusion network with the segmentation model.

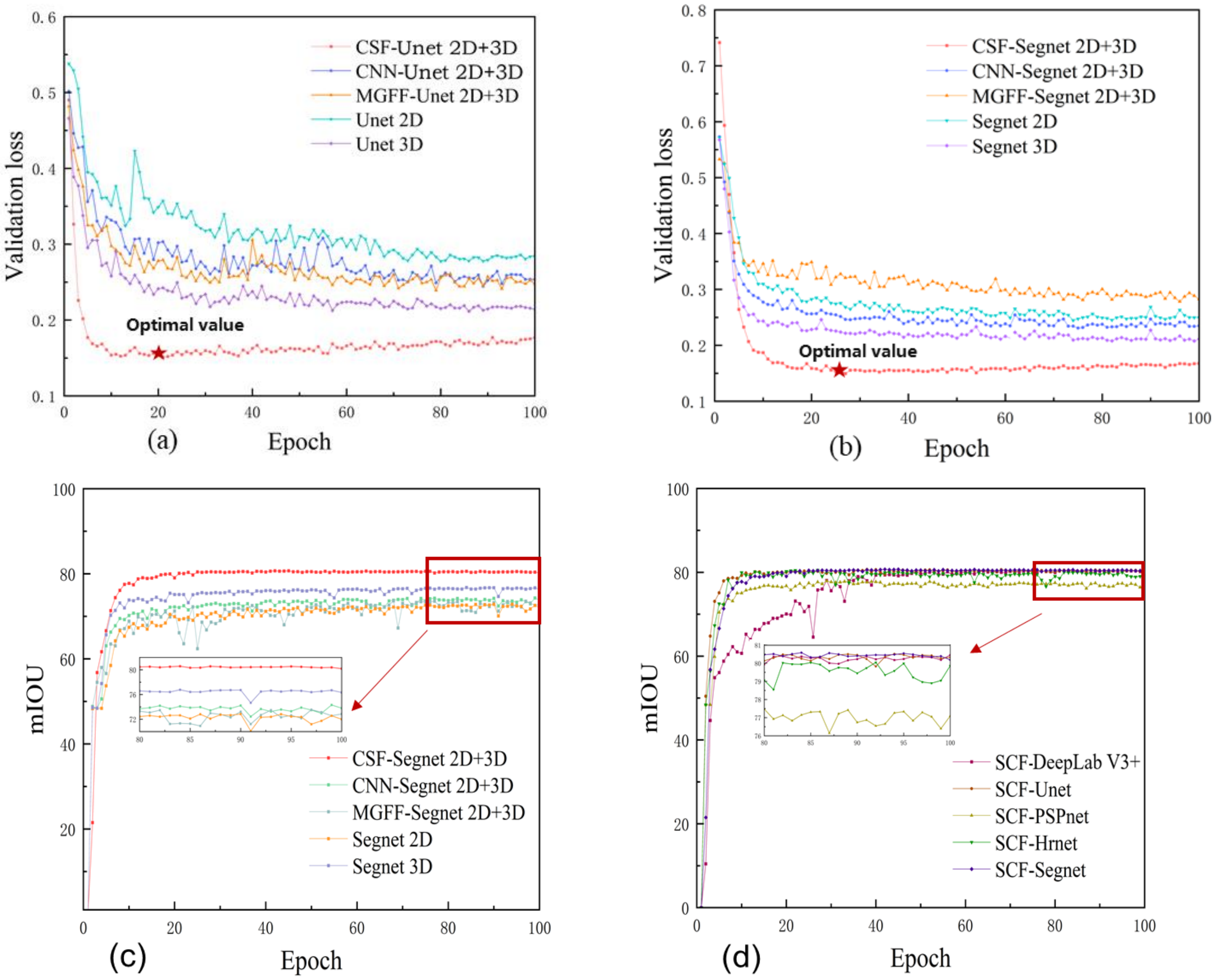


Figure 8 Validation loss and mIOU based on different frameworks and datasets. (a)-(b): loss curves of different datasets using the Unet andSegNet model. (c)-(d): mIOU performance results of different datasets and model.

Table 2 presents the specific performance results of 25 experiments on validation images, illustrating the substantial improvements achieved by CSF-CrackNet. The adaptive fusion strategies result in significant gains in mIOU, precision, and recall, demonstrating the technical superiority of our approach over conventional fusion methods. It can be seen from the table that the performance of different models on the same dataset is quite different. However, CSF-CrackNet has a good performance improvement effect on the original model. For example, the mIOU of most models can be increased to 80%. Compared with the original RGB image, the average increase of mIOU is nearly 10%, and the average increase of mIOU is nearly 5% compared with the original depth image. Other evaluation metrics can also reflect similar results. Compared with other fusion methods, CSF-CrackNet also exhibits better performance. Overall, CSF-CrackNet can be flexibly deployed at the forefront of most semantic segmentation networks, enhancing the performance of segmentation models.

Table 2 Comparison of segmentation results

Framework	Model	mIOU	F1	mAP	Precision	Recall
Deeplab V3+	DeepLab V3+ 2D	67.16%	74.10%	71.93%	83.05%	71.93%
	DeepLab V3+ 3D	73.46%	81.20%	78.11%	88.24%	78.11%
	MGFF-DeepLab V3+ 2D+3D	73.08%	81.10%	78.53%	86.57%	78.53%
	CNN-DeepLab V3+ 2D+3D	73.41%	81.30%	76.02%	88.18%	76.02%
	CSF-DeepLab V3+ 2D+3D	**80.31%**	**89.22%**	**87.15%**	**88.77%**	**87.15%**
Unet	Unet 2D	69.11%	77.00%	72.89%	87.31%	72.89%
	Unet 3D	75.54%	84.00%	82.35%	88.11%	82.35%
	MGFF-Unet 2D+3D	73.64%	82.00%	78.89%	87.27%	78.89%
	CNN-Unet 2D+3D	73.47%	81.80%	78.84%	86.97%	78.84%
	CSF-Unet 2D+3D	**80.50%**	**88.00%**	**86.83%**	**89.97%**	**86.63%**
PSPnet	PSPnet 2D	68.10%	72.90%	83.90%	87.33%	72.90%
	PSPnet 3D	71.69%	80.00%	76.23%	87.33%	76.23%
	MGFF-PSPnet 2D+3D	71.29%	77.90%	77.52%	83.99%	77.52%
	CNN-PSPnet 2D+3D	70.46%	78.30%	75.76%	84.90%	75.76%
	CSF-PSPnet 2D+3D	**77.23%**	**85.00%**	**84.93%**	**86.21%**	**84.93%**
Hrnet	Hrnet 2D	71.04%	80.70%	76.27%	85.50%	76.27%
	Hrnet 3D	75.90%	85.00%	80.93%	88.92%	80.93%
	MGFF-Hrnet 2D+3D	74.04%	83.40%	80.09%	86.25%	80.09%
	CNN-Hrnet 2D+3D	74.03%	82.74%	79.79%	86.69%	79.79%

	CSF-Hrnet 2D+3D	**79.81%**	**87.10%**	**86.48%**	**88.72%**	**86.48%**
Segnet	Segnet 2D	72.98%	81.90%	79.65%	84.65%	79.65%
	Segnet 3D	76.90%	84.50%	83.09%	87.79%	83.09%
	MGFF-Segnet 2D+3D	73.52%	82.20%	79.28%	86.35%	79.28%
	CNN-Segnet 2D+3D	74.40%	82.80%	80.86%	85.92%	80.86%
	CSF-Segnet 2D+3D	**80.51%**	**88.50%**	**87.45%**	**88.76%**	**87.54%**

4.1.3 Visual comparison among different models

Figure 9 shows the segmentation results from different datasets using the Unet model on the test dataset. The pixels of the cracks in the RGB images are similar to the background pixels and are greatly affected by shadows and road attachments. The edge information of the cracks in the depth images is not obvious, and there are many noise points. All these make it difficult for any experienced engineer to obtain a complete image of pavement cracks. For example, in the example of the first row in Figure 9, the fourth column of RGB image segmentation results are affected by shadows, resulting in obvious false-positive errors. However, the depth image will not be affected by illumination, so the depth image segmentation results in this data group are suitable. The segmentation result of the depth image in the fourth column of the sixth row has a false-negative error in the segmentation result of the region in the image's upper left corner due to local subsidence. Segmentation of micro-cracks is challenging due to the unclear 3D characterization, leading to false-negative errors at crack ends in depth image segmentation. Methods like MGFF and CNN partially succeed in fusing RGB and depth images, producing good results in some cases. However, they struggle with discontinuous cracks, lacking robustness to adapt to the unique characteristics of pavement cracks. CSF-CrackNet integrates features from both RGB and depth images, enabling it to effectively address the challenges of crack segmentation in most scenarios. It uses multi-dimensional image fusion and adaptive channel weights to accurately segment crack widths despite distortions caused by water stains, leveraging the complementary strengths of 2D RGB and 3D images. As shown in the fourth row, this approach ensures precise segmentation even under challenging conditions. However, it tends to produce false positive errors in the segmentation of crack intersection points, especially in complex mesh cracks, such as the last row of results in Figure 9.

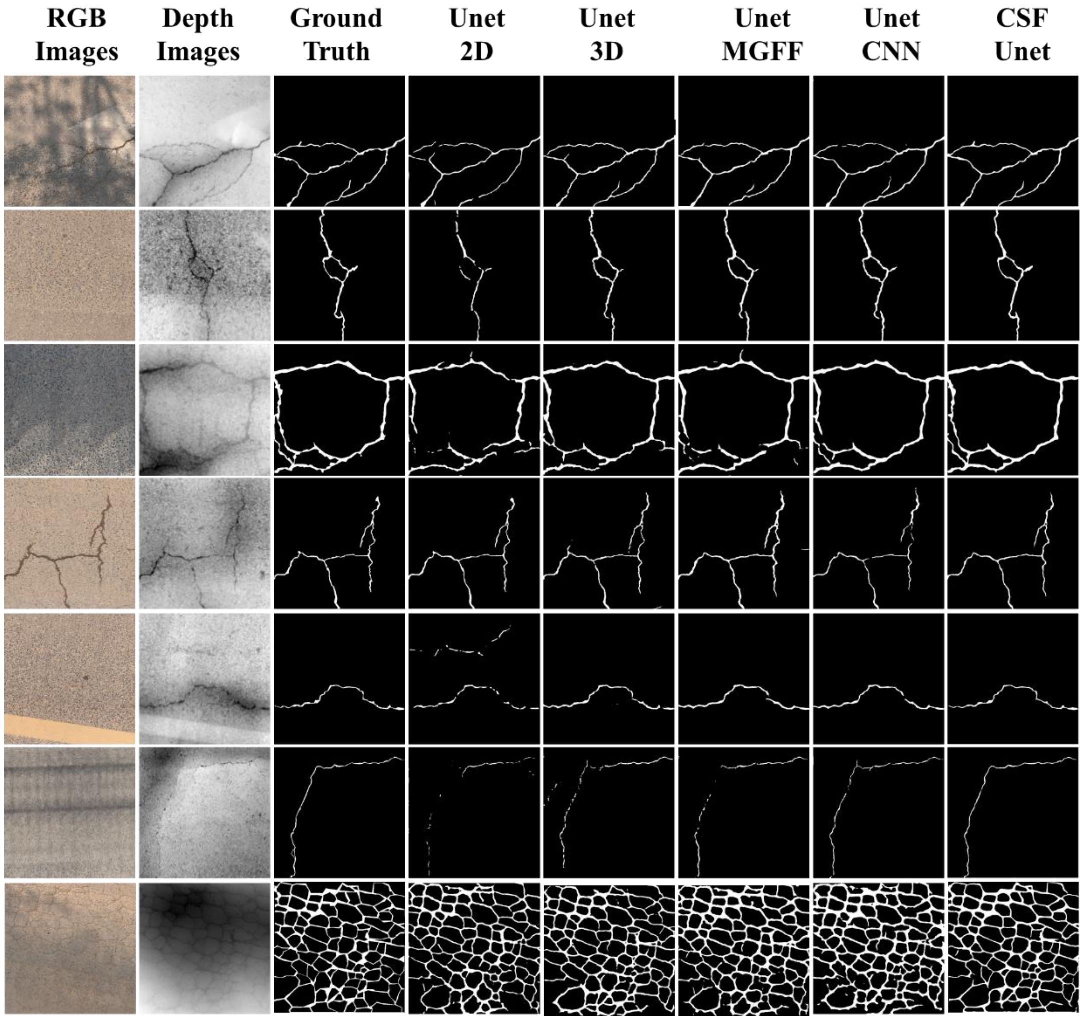


Figure 9. Visual prediction results and comparison between our model and previous models with Unet framework

Figure 10 shows the results of segmentation using CSF-CrackNet based on different frameworks. It can be seen from this figure that our method can better segment the crack pixels from the background pixels. The DeepLab V3 + framework has a good effect on complex fractures because it uses the ASPP network to adapt to the fracture characteristics of different scales. The Unet network's exceptional information extraction capability contributes to superior segmentation results for the crack continuity preservation. Specifically, the Unet framework excels in retaining the continuous characteristics of cracks. Conversely, the Segnet framework employs an index method for up-sampling. In essence, the pooling operation records the position of the value, enabling direct UpPooling with the position recorded position during up-sampling., This approach yields favorable segmentation results for the edge characteristics of the crack. The Hrnet and the Pspnet frameworks exhibit slightly inferior performance compared to other frameworks. However, they still demonstrate improved performance compared to the original network results.

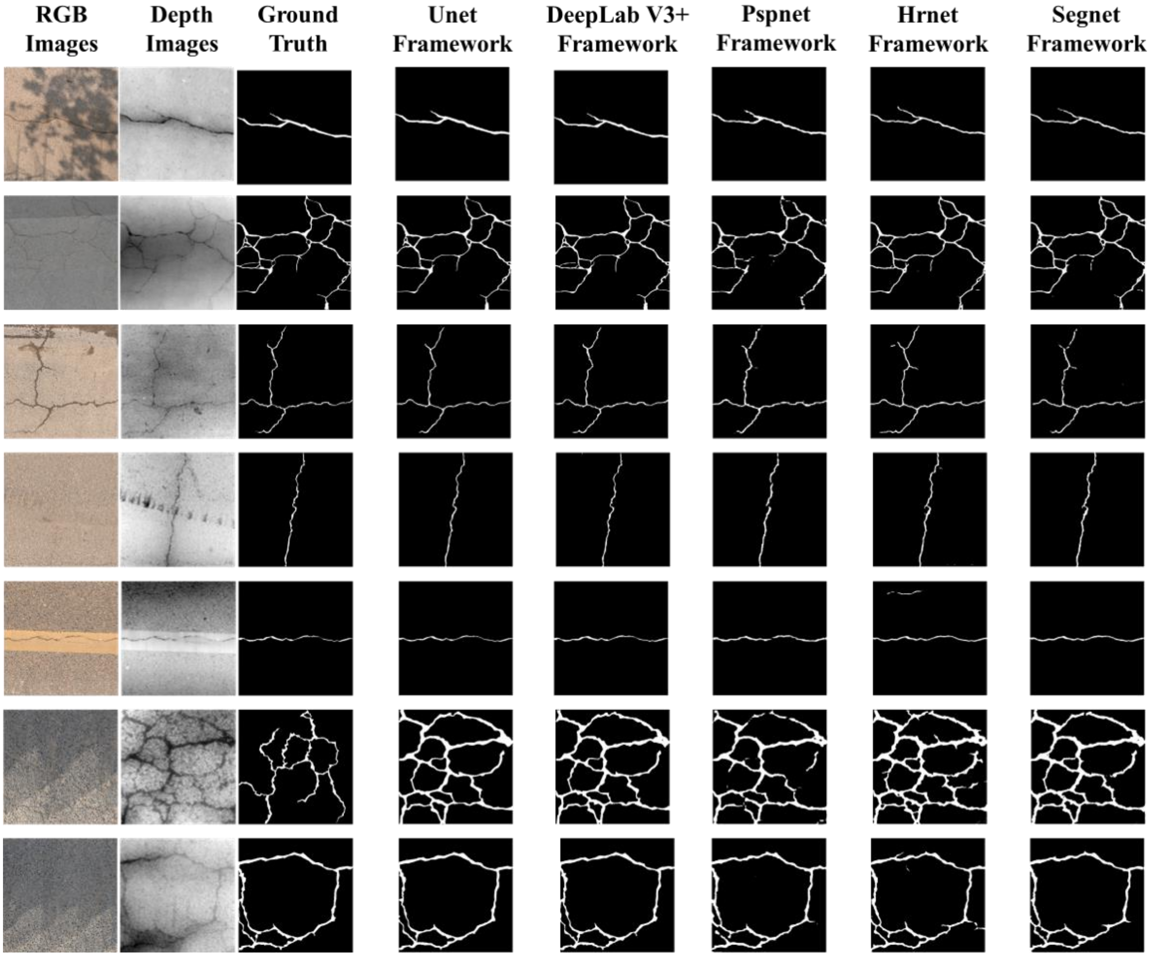


Figure 10. Visual prediction results and comparison of different frameworks using CSF-CrackNet (ours)

4.2 Evaluation of model complexity

The parameters number and processing time for each network are shown in Table

3. The incorporation of CSF-CrackNet does introduce an additional computational load to the operation of the underlying pavement crack segmentation network. It is crucial to highlight that the processing time provided in Table 3 is obtained within a computational environment that includes a personal computer equipped with an NVIDIA GeForce RTX 3090, and the input image pixel size is 512 × 512 (width × height). Various factors, such as image size and computer performance, may influence the processing time. The calculation time of the model is related to the parameters. In contrast, the introduction of CSF-CrackNet leads to a significant increase in model parameters. However, many residual structures are introduced into the model, which makes the calculation speed less affected, and the model can still be flexibly deployed in mobile/low-performance devices. Despite the compromise in fast calculation capability with the deployment of CSF-CrackNet, we deem it worthwhile as it significantly enhances the accuracy of pavement crack semantic segmentation. Additionally, the network structure can be pruned to suit the requirements of practical tasks[44]. In the case described in Section 4.1.3, the CSF-DeepLab V3 + model with the slowest processing speed is taken as an example. The most significant addition in CSF-DeepLab V3+ is the increase in computational complexity, leading to a more than 38% increase in computation time. However, it has also achieved excellent performance. It is worth noting that the increase in computation time is more pronounced when deploying large parameter networks. This is because when the number of basic parameters in the network model is very large, updating each parameter requires computational resources, resulting in a significant increase in computation time. At the same time, many parameters need to be stored in memory for updating during training. Memory limitations can also result in slower computation speeds. In contrast, the CSF-Segnet, which has a smaller number of basic parameters, introduces almost the same number of parameters as the CSF-CrackNet, with only a 9% increase in computation time. It still achieves good pixel-level segmentation performance. Therefore, we believe the additional computation time can be reduced by using more efficient computing

hardware or distributed training methods, further optimizing network performance.

Table 3 Processing time and parameters of models

Framework	Model	Processing Time (ms/Frame)	Parameters
Deeplab V3+	CSF-DeepLab V3+ 2D+3D	69.54	59,623,713
	DeepLab V3+ 3D	50.05	54,708,674
	DeepLab V3+ 2D	50.79	54,708,674
	MGFF-DeepLab V3+ 2D+3D	49.90	54,708,674
	CNN-DeepLab V3+ 2D+3D	49.64	54,708,674
Unet	CSF-Unet 2D+3D	47.55	48,847,905
	Unet 3D	36.95	43,932,866
	Unet 2D	37.31	43,932,866
	MGFF-Unet 2D+3D	33.48	43,932,866
	CNN-Unet 2D+3D	33.36	43,932,866
PSPnet	CSF-PSPnet 2D+3D	53.42	51,621,945
	PSPnet 3D	41.27	46,706,626
	PSPnet 2D	39.98	46,706,626
	MGFF-PSPnet 2D+3D	41.56	46,706,626
	CNN-PSPnet 2D+3D	39.51	46,706,626
Hrnet	CSF-Hrnet 2D+3D	22.21	14,551,831
	Hrnet 3D	19.47	9,636,512
	Hrnet 2D	19.52	9,636,512
	MGFF-Hrnet 2D+3D	19.03	9,636,512
	CNN-Hrnet 2D+3D	19.23	9,636,512
Segnet	CSF-Segnet 2D+3D	30.28	32,240,065
	Segnet 3D	27.73	27,322,178
	Segnet 2D	27.12	27,322,178
	MGFF-Segnet 2D+3D	28.34	27,322,178
	CNN-Segnet 2D+3D	26.43	27,32,2178

5. Discussion

5.1 Ablation experiments

To verify the validity of modules in the CSF-CrackNet, Table 4 shows the different combinations of modules used for the ablation experiments. This section aims to discuss the improvement of the effect of the module rather than explain the characteristics of the semantic segmentation network. Therefore, the experimental results listed in Table 4 are based on the Unet framework. It can be seen that the networks obtained by both modules perform better than combining them individually. Compared with the original Unet, the channel module is the most effective, which can increase the mIOU by about 5 %. The space module also has a positive effect on the improvement of segmentation accuracy. The above conclusions show that each module plays an important role in the fusion process.

Table 4 Results of ablation experiments

	dataset	Channel Module	Spatial Module	mIOU	F1	mAP	Precision	Recall
#1	RGB images	none	none	75.54%	84.00%	82.35%	88.11%	82.35%
#2	Depth images	none	none	69.11%	77.00%	72.89%	87.31%	72.89%
#3	RGB images + Depth images	√	none	79.11%	86.00%	85.55%	88.03%	86.02%
#4	RGB images + Depth images	none	√	78.33%	85.00%	85.01%	87.21%	85.21%
#5	RGB images + Depth images	√	√	**80.50%**	**88.00%**	**86.83%**	**89.97%**	**86.63%**

5.2 Self-adapting channel weight

The weight of channel fusion is the weight of each channel feature map, which controls the fusion degree of different channel feature maps. Since the feature maps of different channels contain different image information, the setting of channel fusion weights is crucial for the final image feature extraction and visual effect. If the quality of the feature map of a channel is poor, the weight of the channel should not be too large. Otherwise, it will affect the final feature extraction effect and image quality. To achieve a better channel fusion effect, it is necessary to adjust and optimize images of different quality. Therefore, the image quality determines the channel fusion weight, which can help us better control the fusion degree of different channel feature maps, improve the accuracy and robustness of image feature extraction, and finally obtain better image quality and crack segmentation effect.

The channel fusion strategy proposed in this paper uses the one-dimensional convolution method described in Section 2.2 to extract channel weights. Figure 11 depicts the channel fusion weights of the proposed method based on the Unet framework. Below the image are the weights of the deep, red, green, and blue channels, respectively. The RGB image uses a linear combination of three colors components to represent the color, and any color is related to these three components. The images obtained in the natural environment are easily affected by natural lighting, occlusion and shadows, and the sensitivity of different color channels to information such as brightness is different. The weight distribution in Figure 11 shows that the blue channel has the best information representation ability in this experimental sample compared with other color channels. In contrast, the red channel has the weakest representation ability. This conclusion can also be seen through the images in Figure 11, especially the information representation ability of the crack edge position. In short, calculated weights are consistent with human perception of the image.

The depth channels of the image shown in Figure 11 (a) and (b) have large weights. This is because there is a shadow in the RGB image of Figure 11 (a). And the depth of the micro-cracks in Fig. 11 (b) is shallow, resulting in no obvious color difference. From a large number of experiments, it seems that images with shadows, water stains, and wheel paths tend to have a higher weight in the depth channel. The depth channel of the images shown in Figure 11 (c) (d) has a small weight. Figure 11 (c) shows that the crack edge distribution is irregular and there is local subsidence in Figure 11 (d), which are the reasons for the low depth weight.

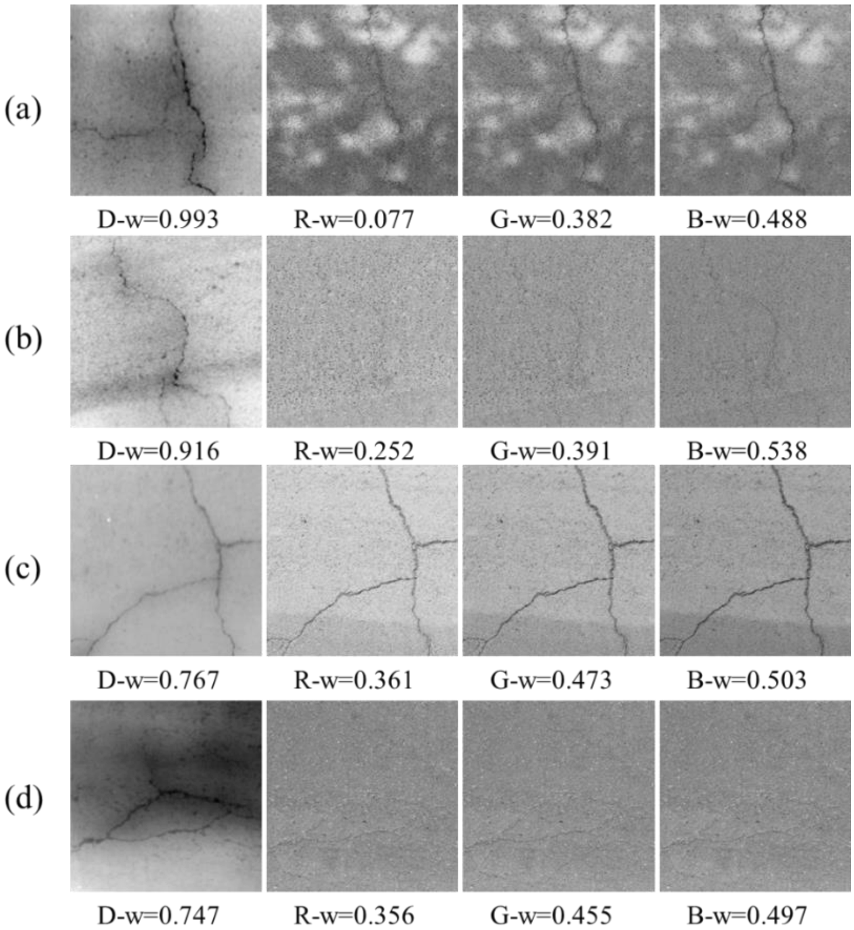


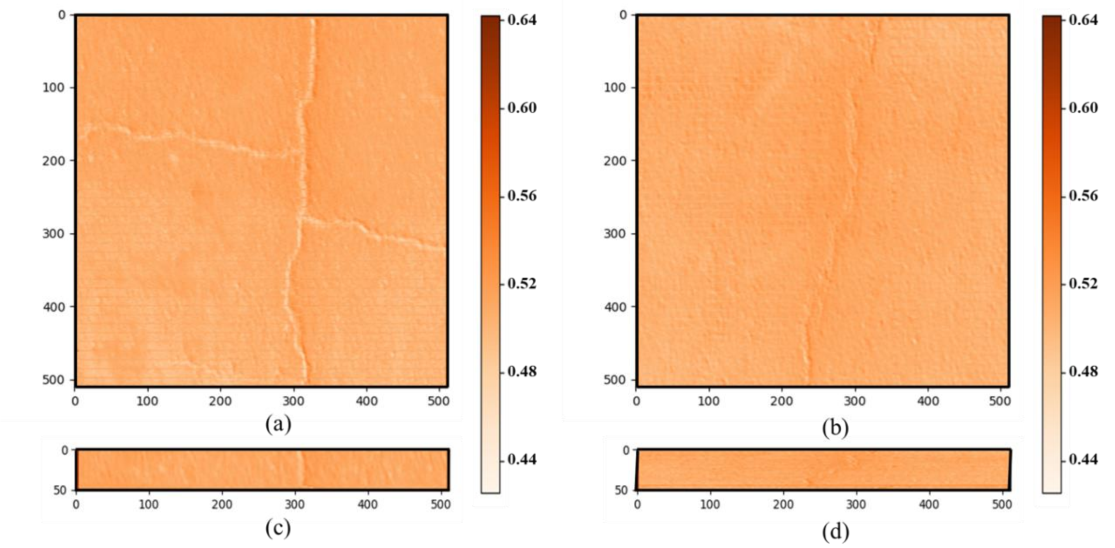
Figure 11. Adaptive channel fusion weight display based on SCF-Unet framework.

5.3 Self-adapting space weight

Spatial fusion can be regarded as a self-adapting spatial region selection mechanism. Not all regions in the image are equally important to the task's contribution. Only task-related regions, particularly in crack segmentation, require attention. By employing self-adapting spatial fusion, the feature information expands its receptive field, thereby strengthening the feature map information.

Figure 12 (a) and (b) show the weight distribution of crack images space fusion in the form of a heat map, and it can be seen that the crack areas, especially the crack edge information, are strengthened. In the adaptive calculation of the spatial feature enhancement matrix, one row of eigenvalues with a larger value is distributed every ten rows, presented in Figure 12 (a)(b) as multiple evenly distributed darker horizontal lines.

To explore the causes of this phenomenon, this paper selects the rows with larger values and draws them as a heat map to get Figure 12 (c)(d). From the figure it is apparent that the image in Figure 12 (a)(b) exhibit similarities to the vertical reduction of the image in Figure 12 (c)(d). This part of the characteristic value still stores the information about the crack, and the larger weight can better retain the original information about the crack. Therefore, the reason for considering this phenomenon is that the original morphological information of cracks in the previous linear pooling layer is retained by uniform sampling.



(a)

(b)

(c)

(d)

Figure 12. Adaptive spatial fusion weight display based on SCF-Unet framework. (a)-(b): Weight distribution of crack images space fusion in the form of a heat map. (c)-(d): The high weight position is displayed in the form of heat map.

5.4 Limitations and future work

This paper introduces CSF-CrackNet, a groundbreaking multidimensional image analysis method that innovatively utilizes channel and spatial fusion. The model's ability to adaptively integrate RGB and depth data represents a significant advancement in pavement crack detection technology the results are better than those of other methods on the dataset used in this paper. However, there are still some problems that are worthy of further study.

(1) Data collection: This paper generates a 3D point cloud model using Structure from Motion (SfM) from multi-view images. Then, it is converted to an orthographic image. Although this method can obtain depth images at a lower cost, it requires a lot of computing resources and time to process the point cloud conversion. In addition, the noise and error introduced in data generation are unavoidable. Improving data accuracy is one of the important ways to improve computing performance. Although 3D laser imaging can also be used to generate depth images quickly and efficiently, it also faces the problem of high cost.

(2) Based on the operation principle of neural networks, this method should be able to be deployed in the front end of semantic segmentation and target detection networks. However, subject to datasets and detection methods, this paper does not try to combine the network with target detection network frameworks such as Yolo. Future research will take this issue into consideration.

6. Conclusions

To improve the accuracy and robustness of pavement crack segmentation, this paper proposes an adaptive fusion method of pavement multi-dimensional images based on channel and space modules, which can be easily and quickly deployed in the front end of the most common semantic segmentation network. CSF-CrackNet is then compared with MGFF and CNN regarding numerical evaluation and visualization results. Finally, we discuss the validity and enhancement mechanism of the model through weight analysis of feature maps. The main contributions and findings of the work can be summarized as follows:

(1) We created a comprehensive pavement crack dataset using Structure from Motion (SfM), which includes various crack forms and complex scenarios. This dataset provides a robust foundation for evaluating crack segmentation networks and ensures the method’s applicability to real-world conditions.

(2) CSF-CrackNet employs an adaptive 2D-3D image fusion mechanism that integrates

the rich color information from RGB images with the structural details from depth images. Advanced channel and spatial modules autonomously learn and apply optimal weights for different image channels and spatial regions. This dynamic adjustment addresses issues like shadows and varying lighting in RGB images, as well as fine detail loss in depth images, by emphasizing informative features and suppressing problematic information from each source. By combining the complementary strengths of RGB and depth data, CSF-CrackNet effectively mitigates environmental noise and enhances segmentation precision. This ensures robust segmentation performance across diverse real-world scenarios by leveraging both the visual details from RGB images and the spatial information from depth images.

(3) Advanced Modules for Robust Feature Extraction and Real-world Performance: CSF-CrackNet incorporates several innovative modules, including the improved Receptive Field Block (RFB), Strip Pooling, one-dimensional convolution and linear fully connected layers, and Diversely Connected Multi-Scale Convolution Block. These modules enhance feature extraction, spatial weighting, and channel fusion, contributing to the model's superior performance. The improved RFB enhances capture of fine details, Strip Pooling improves spatial context integration, the one-dimensional convolution and linear fully connected layers optimize channel fusion, and the Diversely Connected Multi-Scale Convolution Block ensures robust feature abstraction across scales. These enhancements enable CSF-CrackNet to maintain high accuracy under diverse and challenging real-world conditions, such as varying illumination, shadows, and road surface irregularities. The model's robustness makes it particularly suitable for practical engineering applications, ensuring its utility in real-world pavement crack detection tasks.

(4) CSF-CrackNet is designed to seamlessly integrate with a range of established semantic segmentation networks, including DeepLab V3+, Unet, PSPNet, HRNet, and SegNet. Experimental results demonstrate significant performance

improvements across these networks, with the adaptable fusion strategies of CSF-CrackNet enhancing the mIOU of most models to around 80%. This reflects an average increase of nearly 10% compared to the original RGB image and about 5% compared to the original depth image. Other evaluation metrics have also shown substantial improvement. Furthermore, CSF-CrackNet's design ensures it can be flexibly deployed in the front end of most common semantic segmentation networks, highlighting its strong potential for broad and effective integration across diverse architectures.

(5) The study provided visual and analytical evidence of CSF-CrackNet’s effectiveness through channel and spatial weight outputs. These weights align with human intuitive assessments, with regions of clearer crack texture receiving higher weights. This alignment demonstrates the model's ability to accurately prioritize critical image features, thus enhancing information fidelity and segmentation accuracy. The effectiveness of the adaptive weights in CSF-CrackNet highlights its capability to dynamically respond to varying image conditions, ensuring superior segmentation outcomes.

There are still many challenges, including the computational complexity of the 3D point cloud model generation and the slower calculation speed introduced by additional modules. The neural network's operational mechanism requires further analysis, highlighting the need for a more thorough understanding and potential application of network self-regulating feedback for semi-supervised learning. Additionally, the paper suggests the unexplored integration of the method with target detection network frameworks, such as Yolo, presenting promising further research.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 52078049, Grant 52378431 and Grant 52408454, in part by the Fundamental Research Funds for the Central Universities, CHD under Grant

300102210302 and Grant 300102210118 and in part by the 111 Project of Low Carbon Smart Road Infrastructure Construction and Maintenance Discipline Innovation and Talent Introduction Base of Shaanxi Province.

References

[1] J.X. Wang, J.H. Gao, Z.Y. Wang, W. Lv, Research on the Application of Deep Learning Algorithm in Big Data Image Classification, World Conference on Intelligent and 3-D Technologies (WCI3DT), Electr Network, 2022, pp. 459-469, https://doi.org/10.1007/978-981-19-7184-6_38.

[2] Z.C. Xu, Z. Dai, Z.Y. Sun, C. Zuo, H.S. Song, C.W. Yuan, Enhancing Pavement Distress Detection Using a Morphological Constraints-Based Data Augmentation Method, Coatings 13 (2023) 764, https://doi.org/10.3390/coatings13040764.

[3] R.Q. Ren, P.X. Shi, P.J. Jia, X.Y. Xu, A Semi-Supervised Learning Approach for Pixel-Level Pavement Anomaly Detection, Ieee Transactions on Intelligent Transportation Systems 24 (2023) 10099-10107, https://doi.org/10.1109/tits.2023.3267433.

[4] H. Zhang, A.A. Zhang, A.Z. He, Z.S. Dong, Y. Liu, Pixel-level detection of multiple pavement distresses and surface design features with ShuttleNetV2, Structural Health Monitoring-an International Journal 23 (2023) 1263-1279, https://doi.org/10.1177/14759217231183656.

[5] Z. Tong, T. Ma, W.G. Zhang, J. Huyan, Evidential transformer for pavement distress segmentation, Computer-Aided Civil and Infrastructure Engineering 38 (2023) 2317-2338, https://doi.org/10.1111/mice.13018.

[6] H.L. Lin, GoogleNet transfer learning with improved gorilla optimized kernel extreme learning machine for accurate detection of asphalt pavement cracks, Structural Health Monitoring-an International Journal 23 (2024) 2853-2868, https://doi.org/10.1177/14759217231215419.

[7] A. Zhang, K.C.P. Wang, B.X. Li, E.H. Yang, X.X. Dai, Y. Peng, Y. Fei, Y. Liu, J.Q. Li, C. Chen, Automated Pixel-Level Pavement Crack Detection on 3D Asphalt Surfaces Using a Deep-Learning Network, Computer-Aided Civil and Infrastructure Engineering 32 (2017) 805-819, https://doi.org/10.1111/mice.12297.

[8] A. Zhang, K.C.P. Wang, Y. Fei, Y. Liu, S.Y. Tao, C. Chen, J.Q. Li, B.X. Li, Deep Learning-Based Fully Automated Pavement Crack Detection on 3D Asphalt Surfaces with an Improved CrackNet, Journal of Computing in Civil Engineering 32 (2018) 04018041, https://doi.org/10.1061/(asce)cp.1943-5487.0000775.

[9] Y. Fei, K.C.P. Wang, A. Zhang, C. Chen, J.Q. Li, Y. Liu, G.W. Yang, B.X. Li, Pixel-Level Cracking Detection on 3D Asphalt Pavement Images Through Deep-Learning- Based CrackNet-V, Ieee Transactions on Intelligent Transportation Systems 21 (2020) 273-284, https://doi.org/10.1109/tits.2019.2891167.

[10] Y. Liu, G. Yang, K.C.P. Wang, G. Wang, J. Li, T. Nantung, Automatic detection of deteriorated inverted-T patching using 3D laser imaging system based on a true story indiana, Intelligent Transportation Infrastructure 1 (2022), https://doi.org/10.1093/iti/liac011.

[11] L.L. Liu, M.L. Chen, M.L. Xu, X.L. Li, Two-stream network for infrared and visible images fusion,

Neurocomputing 460 (2021) 50-58, https://doi.org/10.1016/j.neucom.2021.05.034.

[12] K.R. Prabhakar, V.S. Srikar, R.V. Babu, Ieee, DeepFuse: A Deep Unsupervised Approach for Exposure Fusion with Extreme Exposure Image Pairs, 16th IEEE International Conference on Computer Vision (ICCV), Ieee, Venice, ITALY, 2017, pp. 4724-4732, https://doi.org/10.1109/iccv.2017.505.

[13] H. Li, X.J. Wu, T.S. Durrani, Infrared and visible image fusion with ResNet and zero-phase component analysis, Infrared Physics & Technology 102 (2019) 103039, https://doi.org/10.1016/j.infrared.2019.103039.

[14] J.C. Guan, X. Yang, L. Ding, X.Y. Cheng, V.C.S. Lee, C. Jin, Automated pixel-level pavement distress detection based on stereo vision and deep learning, Automation in Construction 129 (2021) 103788, https://doi.org/10.1016/j.autcon.2021.103788.

[15] D.P. Bavirisetti, G. Xiao, J.H. Zhao, R. Dhuli, G. Liu, Multi-scale Guided Image and Video Fusion: A Fast and Efficient Approach, Circuits Systems and Signal Processing 38 (2019) 5576-5605, https://doi.org/10.1007/s00034-019-01131-z.

[16] R. Heideklang, P. Shokouhi, Multi-sensor image fusion at signal level for improved near-surface crack detection, Ndt & E International 71 (2015) 16-22, https://doi.org/10.1016/j.ndteint.2014.12.008.

[17] G.H. Beckman, D. Polyzois, Y.J. Cha, Deep learning-based automatic volumetric damage quantification using depth camera, Automation in Construction 99 (2019) 114-124, https://doi.org/10.1016/j.autcon.2018.12.006.

[18] W.D. Zhang, L. Zhou, P.X. Zhuang, G.H. Li, X.P. Pan, W.Y. Zhao, C.Y. Li, Underwater Image Enhancement via Weighted Wavelet Visual Perception Fusion, Ieee Transactions on Circuits and Systems for Video Technology 34 (2024) 2469-2483, https://doi.org/10.1109/tcsvt.2023.3299314.

[19] E. Mouaddib, A. Pamart, M. Pierrot-Deseilligny, D. Girardeau-Montaut, 2D/3D data fusion for the comparative analysis of the vaults of Notre-Dame de Paris before and after the fire, Journal of Cultural Heritage 65 (2024) 221-231, https://doi.org/10.1016/j.culher.2023.06.012.

[20] P.G. Li, B. Zhou, C. Wang, G.Z. Hu, Y. Yan, R.X. Guo, H.T. Xia, CNN-based pavement defects detection using grey and depth images, Automation in Construction 158 (2024) 105192, https://doi.org/10.1016/j.autcon.2023.105192.

[21] Y. Liu, X. Chen, J. Cheng, H. Peng, Z.F. Wang, Infrared and visible image fusion with convolutional neural networks, International Journal of Wavelets Multiresolution and Information Processing 16 (2018) 1850018, https://doi.org/10.1142/s0219691318500182.

[22] L. Zhao, H. Zhou, X.G. Zhu, X. Song, H.S. Li, W.B. Tao, LIF-Seg: LiDAR and Camera Image Fusion for 3D LiDAR Semantic Segmentation, Ieee Transactions on Multimedia 26 (2024) 1158-1168, https://doi.org/10.1109/tmm.2023.3277281.

[23] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, Arxiv (2015), http://arxiv.org/abs/1409.1556.

[24] A.Z. He, Z.S. Dong, H. Zhang, A.A. Zhang, S. Qiu, Y. Liu, K.C.P. Wang, Z.H. Lin, Automated Pixel-Level Detection of Expansion Joints on Asphalt Pavement Using a Deep-Learning-Based Approach, Structural Control & Health Monitoring 2023 (2023) 15, https://doi.org/10.1155/2023/7552337.

[25] S.T. Liu, D. Huang, Y.H. Wang, Receptive Field Block Net for Accurate and Fast Object Detection, 15th European Conference on Computer Vision (ECCV), Munich, GERMANY, 2018, pp. 404-419, https://doi.org/10.1007/978-3-030-01252-6_24.

[26] S.H. Huang, Z.C. Lu, R. Cheng, C. He, Ieee, FaPN: Feature-aligned Pyramid Network for Dense Image Prediction, 18th IEEE/CVF International Conference on Computer Vision (ICCV), Electr Network, 2021, pp. 844-853, https://doi.org/10.1109/iccv48922.2021.00090.

[27] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional Networks for Biomedical Image Segmentation, 18th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Munich, GERMANY, 2015, pp. 234-241, https://doi.org/10.1007/978-3-319-24574-4_28.

[28] J. Hyun, H. Seong, E. Kim, Universal pooling-A new pooling method for convolutional neural networks, Expert Systems with Applications 180 (2021) 115084, https://doi.org/10.1016/j.eswa.2021.115084.

[29] T.Y. Lin, P. Dollár, R. Girshick, K.M. He, B. Hariharan, S. Belongie, Ieee, Feature Pyramid Networks for Object Detection, 30th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Ieee, Honolulu, HI, 2017, pp. 936-944, https://doi.org/10.1109/cvpr.2017.106.

[30] Q.B. Hou, L. Zhang, M.M. Cheng, J.S. Feng, Ieee, Strip Pooling: Rethinking Spatial Pooling for Scene Parsing, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Ieee Computer Soc, Electr Network, 2020, pp. 4002-4011, https://doi.org/10.1109/cvpr42600.2020.00406.

[31] I. Shiri, M. Amini, F. Yousefirizi, A.V. Sadr, G. Hajianfar, Y. Salimi, Z. Mansouri, E. Jenabi, M. Maghsudi, I. Mainta, M. Becker, A. Rahmim, H. Zaidi, Information fusion for fully automated segmentation of head and neck tumors from PET and CT images, Medical Physics 51 (2023) 319-333, https://doi.org/10.1002/mp.16615.

[32] M. Haribabu, V. Guruviah, Enhanced multimodal medical image fusion based on Pythagorean fuzzy set: an innovative approach, Scientific Reports 13 (2023) 16726, https://doi.org/10.1038/s41598-023-43873-6.

[33] Y.Z. He, Y.X. Wang, F.W. Wu, R.Z. Yang, P. Wang, S.B. She, D.T. Ren, Temperature monitoring of vehicle brake drum based on dual light fusion and deep learning, Infrared Physics & Technology 133 (2023) 104823, https://doi.org/10.1016/j.infrared.2023.104823.

[34] Q. Jin, S.Q. Tan, G. Zhang, Z.G. Yang, Y.J. Wen, H.S. Xiao, X. Wu, Visible and Infrared Image Fusion of Forest Fire Scenes Based on Generative Adversarial Networks with Multi-Classification and Multi-Level Constraints, Forests 14 (2023) 1952, https://doi.org/10.3390/f14101952.

[35] J. Xie, W. Li, M. Wang, Research on pavement crack detection algorithm in complex background, 2023 IEEE International Conference on Control, Electronics and Computer Technology (ICCECT) (2023) 1102-1106, https://doi.org/10.1109/iccect57938.2023.10140681.

[36] J.Y. Yuan, S. Li, OMOFuse: An Optimized Dual-Attention Mechanism Model for Infrared and Visible Image Fusion, Mathematics 11 (2023) 4902, https://doi.org/10.3390/math11244902.

[37] B.C. Russell, A. Torralba, K.P. Murphy, W.T. Freeman, LabelMe: a database and Web-based tool for image annotation, International Journal of Computer Vision 77 (2008) 157-173, https://doi.org/10.1007/s11263-007-0090-8.

[38] M.J. Li, Y.B. Dong, X.L. Wang, Research on Image Fusion based on Pyramid Decomposition, 3rd International Conference on Energy, Environment and Sustainable Development (EESD 2013), Trans Tech Publications Ltd, Shanghai, PEOPLES R CHINA, 2013, pp. 2855-+, https://doi.org/10.4028/www.scientific.net/AMR.860-863.2855.

[39] H.F. Li, X.S. Li, Z.T. Yu, C.L. Mao, Multifocus image fusion by combining with mixed-order structure tensors and multiscale neighborhood, Information Sciences 349 (2016) 25-49, https://doi.org/10.1016/j.ins.2016.02.030.

[40] L.-C. Chen, G. Papandreou, F. Schroff, H. Adam, Rethinking Atrous Convolution for Semantic Image Segmentation, arXiv preprint arXiv:1706.05587 (2017), http://arxiv.org/abs/1706.05587.

[41] H.S. Zhao, J.P. Shi, X.J. Qi, X.G. Wang, J.Y. Jia, Ieee, Pyramid Scene Parsing Network, 30th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 2017, pp. 6230-6239, https://doi.org/10.1109/cvpr.2017.660.

[42] K. Sun, B. Xiao, D. Liu, J. Wang, Deep High-Resolution Representation Learning for Human Pose Estimation, Arxiv (2019), http://arxiv.org/abs/1902.09212.

[43] V. Badrinarayanan, A. Kendall, R. Cipolla, SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation, Ieee Transactions on Pattern Analysis and Machine Intelligence 39 (2017) 2481-2495, https://doi.org/10.1109/tpami.2016.2644615.

[44] J. Frankle, M. Carbin, The lottery ticket hypothesis: Finding sparse, trainable neural networks, arXiv preprint arXiv:1803.03635 (2018), http://arxiv.org/abs/1803.03635.