Article

An Optimized Intelligent Segmentation Algorithm for Concrete Cracks Based on Transformer

Tianhao Ye [1], **Min He** [1], **Yexuan Wang** [2], **Jiaying Wang** [3], **Lei Zhu** [2] and **Jie Zhang** [2,*]

[1] School of Civil and Architectural Engineering, Xi'an University of Technology, Xi'an 710048, China; 2230720220@stu.xaut.edu.cn (T.Y.); hem@xaut.edu.cn (M.H.)
[2] School of Computer Science and Engineering, Xi'an University of Technology, Xi'an 710048, China; 2241221121@stu.xaut.edu.cn (Y.W.); leizhu@xaut.edu.cn (L.Z.)
[3] School of Mechanical and Precision Instrument Engineering, Xi'an University of Technology, Xi'an 710048, China; 2230221209@stu.xaut.edu.cn
[*] Correspondence: jiezhang1984@xaut.edu.cn

Abstract: The accurate detection and segmentation of concrete cracks are crucial for maintaining the integrity and safety of infrastructure. Traditional manual inspection methods are often constrained by background complexity and environmental noise, while deep learning models face challenges related to data dependency and poor generalization in complex scenarios. To address these issues, this paper proposes an enhanced transFissNet model that integrates a ResNet101 backbone with Transformer modules for self-attention and multi-scale feature extraction. The model improves the robustness of crack detection under varying lighting conditions and irregular crack morphologies. Experimental results on multiple benchmark datasets demonstrate that transFissNet achieves an accuracy of 96.8%, outperforming existing mainstream methods. The proposed approach provides a reliable and scalable solution for automated crack segmentation and contributes to the advancement of intelligent structural health monitoring.

Keywords: concrete cracks; semantic segmentation; crack detection; convolutional neural network

1. Introduction

Concrete is an indispensable main material in modern infrastructure construction, widely used in structures such as dams, bridges, tunnels, etc. Due to its excellent mechanical properties, concrete has good performance in bearing structural loads and resisting external environmental erosion. However, during long-term service, concrete is almost inevitably prone to cracking because of factors such as external loads, environmental corrosion, and material aging. Cracks are a common type of structural damage that jeopardizes the health of concrete buildings [1,2]. The appearance of cracks not only weakens the overall bearing capacity of concrete structures but also provides channels for the invasion of harmful substances such as water, gas, and chemicals from the outside, accelerating the deterioration process of materials and greatly threatening the service life and structural safety of infrastructure. Therefore, how to accurately and quickly detect and control the propagation of concrete cracks is a key issue to ensure structural safety and extend its service life.

Currently, the detection of concrete cracks is primarily dependent on manual inspection, but with the scaling and complication of infrastructure, its limitations are becoming increasingly prominent [3,4]. Manual detection is costly and inefficient, making it difficult

to adapt to large-scale applications. The detection results are influenced by the subjective judgment of operators, making them susceptible to errors due to varying levels of experience. In complex or hazardous environments, the difficulty and risk of manual detection significantly increase as well [5,6]. Therefore, it is imperative to develop automated and intelligent crack detection methods to replace conventional manual approaches [7–9].

In this context, automated crack detection technology utilizing image processing has increasingly become a focus of research. In recent years, the breakthrough of crack detection mainly depends on the update and development of deep learning technology. Notably, convolutional neural networks (CNNs) in deep learning have been widely applied in various image recognition and segmentation tasks and have shown significant advantages in dealing with complex pattern recognition problems [10–12]. In the field of concrete crack detection, deep learning methods can automatically identify crack areas in complex backgrounds through their powerful feature extraction capabilities, which effectively improves the accuracy and efficiency of crack detection compared with traditional artificial feature extraction and avoids the limitation of the extraction process [13,14]. However, existing deep learning-based detection methods still face many challenges in practical applications [15,16]. First, the generalization ability of the model still needs to be improved, and the performance of existing methods is not stable enough when facing different lighting conditions, complex backgrounds, and a variety of crack morphologies [17]. Second, the dependency on large-scale datasets leads to poor performance of the model in situations where data are scarce or incompletely labeled, which affects its practical application promotion.

To solve these problems, this paper proposes an improved intelligent segmentation network for concrete cracks—transFissNet. Based on the ResNet101 backbone network, this model innovatively introduces Transformer modules [18], aiming to better capture the global feature correlations of cracks through the self-attention mechanism. Unlike traditional convolution operations, the self-attention mechanism of Transformers can not only establish dependencies between pixel points on a global scale but also effectively handle segmentation problems under diverse crack morphologies and complex backgrounds. Through this global feature modeling, transFissNet can maintain stronger robustness and generalization ability when facing complex scenarios such as changes in lighting, noise interference, and irregular crack shapes [19]. Additionally, transFissNet combines a multi-scale feature fusion strategy, which further enhances the model's ability to capture crack details. Since cracks often have different scales and complex morphologies, simple feature extraction methods are often insufficient to characterize their features fully. The multi-scale feature fusion strategy comprehensively processes features at different levels to enable the model to not only extract global contextual information but also precisely segment the detailed parts of cracks [20]. This fusion of multi-scale features, combined with the global feature modeling capability of Transformers, enables transFissNet to demonstrate outstanding performance in crack segmentation tasks.

By embedding Transformer modules, our network is able to process global features of images in parallel and capture global correlations of cracks in complex backgrounds without relying on sequence order. The combination of global feature modeling and multi-scale feature fusion not only improves the accuracy of crack segmentation but also effectively enhances the efficiency of crack detection.

The significant contributions of this research can be summarized as follows:

- An improved crack segmentation model transFissNet based on ResNet101 and Transformer has been proposed, which innovatively introduces a self-attention mechanism, effectively enhancing the global feature capture capability of crack detection.

- The multi-scale feature fusion strategy has been introduced, which can not only accurately segment cracks but also intelligently repair crack areas, greatly enhancing the robustness and accuracy of crack recognition and repair.
- A crack intelligent recognition and repair system has been proposed, which can automatically detect and repair cracks, providing a complete automated solution from recognition to repair, significantly improving the efficiency of crack processing and maintenance.

The rest of this article is structured as follows. The second part investigates and analyzes the current research status and application of image segmentation, transfer learning, and deep learning in the field of concrete crack detection and segmentation. In addition, the principles and applications of the self-attention mechanism are studied and elaborated. The third part elaborates on our work framework and model architecture in detail. The fourth part presents experimental details and results by comparing them with other neural network models. Finally, we summarize the entire text and provide a summary and outlook for future work.

2. Related Work

2.1. Image Segmentation and Detection Methods

Image segmentation has significant advantages in the localization and classification of multiple objects in images. Compared with traditional segmentation methods that rely on low-level visual information, semantic segmentation based on deep learning utilizes high-order visual information, demonstrating stronger generalization ability and robustness [21–23]. High-order visual methods can extract image features and capture contextual relationships between pixels and their surroundings using deep neural network technology, thereby reducing ambiguity in segmentation results [24,25]. Dong et al. [26] proposed a crack detection method based on an improved ResNet-14 and U-shaped Swin-Unet network (RS-Unet) for concrete crack detection, which demonstrated excellent anti-interference and accuracy for small cracks under noise interference. Xu et al. [27] constructed a benchmark dataset for concrete crack segmentation using an optimized DeepLabv3+ segmentation algorithm, combined with a mixture of high-quality and low-quality samples and trained it on ResNet101. Although this method has high computational complexity, its effectiveness in handling small targets and boundary details remains limited. Meng et al. [28] explored an end-to-end crack segmentation model utilizing ResNet101, employing manually annotated and expanded real-world crack datasets for evaluation. Despite its effectiveness, this model's accuracy improvement was limited.

While image segmentation-based crack detection methods are relatively accurate in crack localization, they still exhibit shortcomings in micro-crack classification. Object detection algorithms, on the other hand, have demonstrated excellent performance in crack recognition and classification, with commonly used methods including YOLO [29,30], SSD [31], and R-CNN [32]. However, the majority of existing studies concentrate on either single-task recognition or sequential dual-task recognition, leaving multi-task integration for crack detection largely underexplored. Xu et al. [33] introduced a bilateral segmentation network, extended convolution, and a pyramid pooling module based on YOLOv5 (v6.2), proposing the YOLOv5-IDS model, which achieved a mAP@0.5 of 84.33% and a mIoU of 94.78%, significantly improving crack detection accuracy and efficiency.

Despite the strong performance of these object detection algorithms in crack recognition, challenges remain when addressing large aspect ratios, structural overlap, and pronounced directional features of cracks. Yu et al. [34] improved YOLOv5 and proposed R-YOLOv5, enhancing crack detection efficiency by introducing angle regression variables. A novel loss function was proposed, along with the integration of PSA Neck and

ECA-Layer modules. Upon training with 1628 crack images, the model attained a mean Average Precision (mAP) of 94.03% at an IoU of 0.5, though its robustness under varying lighting conditions and noise environments still requires further validation. Similarly, Liu et al. [35] employed Eigen-CAM for the visual interpretation of crack classification based on YOLOv5, revealing performance differences of CNN models under different crack shapes and defect types. While object detection algorithms have shown excellent performance in crack recognition, challenges remain in addressing the diversity of crack shapes, robustness in noisy environments, and the detection of small cracks [36–38].

In summary, classical models such as FCN and UNet offer fast convergence and stable training but may struggle with precise boundary localization [39]. PSPNet and DeepLabV3 leverage multi-scale context information, yet their performance can degrade in low-contrast crack images. Transformer-based models such as TransUNet improve global feature modeling but often suffer from high computational costs. In contrast, our proposed transFissNet combines the strengths of CNN and Transformer architectures, achieving a better trade-off between segmentation accuracy, robustness, and model efficiency.

2.2. Transfer Learning

Transfer learning has made significant progress in the field of concrete crack detection. With the advancement of big data and deep learning technologies, transfer learning has gained significant attention as an efficient method to leverage existing knowledge for addressing new challenges. In crack detection tasks, transfer learning can utilize pre-trained models from other related tasks to extract and transfer useful features, thereby reducing the reliance on large amounts of annotated data and improving detection accuracy and efficiency. Researchers have explored various transfer learning strategies, including fine-tuning pre-trained models, domain adaptation, multi-task learning, etc., to adapt to different detection scenarios and needs.

Nowadays, the combination of concrete crack detection and transfer learning is widely studied and applied. For example, Su et al. [40] combined deep convolutional neural networks (CNNs) and transfer learning techniques to optimize crack detection performance and improve detection accuracy. Wang et al. [41] proposed a simplified real-time detection model based on the Transformer architecture. Through the integration of a receptive field attention module and a feature allocation mechanism, the model demonstrated improved accuracy and efficiency in crack detection. In addition, Fang et al. [42] proposed an external attention-based TransUNet architecture combined with a label expansion strategy, achieving promising results in crack detection tasks. The model (RAI-DETR) has been experimentally validated for its effectiveness in detecting concrete cracks. In addition, Philip et al. [43] studied the application of transfer learning in crack detection of concrete walls and found that ResNet50 performed the best in classifying crack images, with high accuracy and shorter training time. Jie et al. [44] introduced the Transformer module to extract global features of images and used U-Net for detail recovery and precise segmentation, significantly reducing training time and the need for large-scale datasets, demonstrating the potential of transfer learning in structural health monitoring.

Although transfer learning has shown good results in crack detection, it still faces some challenges. First, disparities between the source domain and the target domain can result in suboptimal transfer performance, especially when the concrete surfaces and crack types vary greatly in different environments. Second, label inconsistency can adversely impact model performance, as deviations in labeled data during practical applications increase the complexity of training and evaluation processes. Furthermore, domain adaptation techniques are still in the developmental stage, and how to effectively transfer knowledge remains a key issue. Insufficient data on rare crack types further limits the performance of

the model, and the complexity of transfer learning models may also affect the inference efficiency in real-time applications. Therefore, further research is required to address these challenges to enhance the practicality and effectiveness of transfer learning in the detection of concrete cracks.

2.3. Attention Mechanism

Attention mechanism is a technique applied to neural networks that can automatically learn key information from data, enhance attention to key parts, and thus improve model performance. This mechanism is extensively applied in the domains of natural language processing and image recognition and has shown significant effects in concrete crack recognition. The attention mechanism technique can improve the model's ability to focus on the crack area, thereby enhancing the accuracy and robustness of crack recognition.

Xu et al. [45] proposed an enhanced Mask R-CNN model aimed at automating the detection and segmentation of defects on tunnel surfaces. By integrating a self-attention module with a rotatable variable window, the model enhances the ability to capture the rotational features of cracks, compensating for the deficiencies in traditional bridge crack feature extraction and significantly improving the crack segmentation effect. Ranyal et al. [46] integrated the attention mechanism into the RetinaNet model to meet the requirements of lightweight and high inference speed when obtaining road crack images from vehicle-mounted cameras. This method improves the accuracy of crack localization and detection by dynamically focusing on crack-related features and suppressing irrelevant information, especially when dealing with complex crack patterns [47]. After incorporating the spatial attention mechanism, the accuracy of the model in multi-scale detection is significantly improved, especially when dealing with cracks of different sizes and complexities. However, in practical applications, attention mechanisms still need to focus on multi-scale and long-range dependency issues. Concrete cracks often have different scales in images, ranging from small cracks to large-scale cracks, so the model needs to be able to process multi-scale information to capture features comprehensively. In addition, cracks may span large areas, and traditional convolution operations have limitations in capturing long-range dependencies. Although attention mechanisms have advantages, it is necessary to ensure that their design can effectively process this long-distance information to utilize their effectiveness fully.

3. Methodology

The research framework of this paper is illustrated in Figure 1, which primarily comprises two main functions: crack detection and crack segmentation. First, the original image of concrete cracks is input into this integrated model. If cracks are detected, the center coordinates of the cracks will be located, and the edge positions and outer rectangular box areas of the cracks will be analyzed. At the same time, rectangular boxes will be used for annotation. This model is based on transfer learning to segment cracks. Subsequently, adjust the central coordinates and the area results derived from the aforementioned dual tasks. If the coordinates of the two-match the area, the region that is connected will be identified as a solid crack; in contrast, it will be deemed noise, thereby accomplishing the overall segmentation of cracks within concrete.
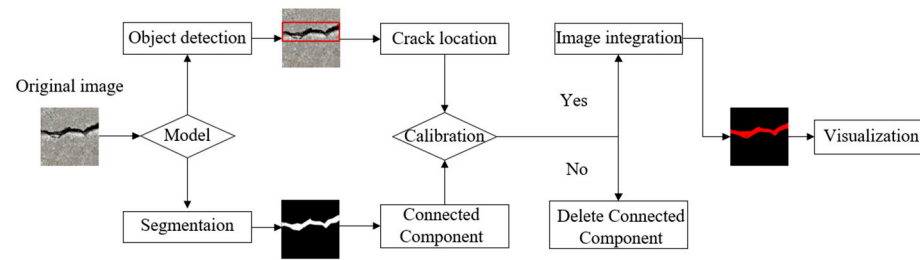
Figure 1. Method framework.

3.1. Overview of ResNet101

ResNet101 is a deep residual network composed of 101 layers with identity shortcut connections, which effectively alleviates the degradation problem in deep networks. Its architecture allows for stable gradient propagation during training and facilitates the extraction of rich and multi-scale semantic features from complex images. Considering the morphological diversity and edge ambiguity of concrete cracks, we adopt ResNet101 as the backbone network to ensure sufficient feature depth and robustness in representation. As illustrated in Figure 2, the network primarily consists of numerous residual blocks, which are the core components of the ResNet architecture. First, the original image is fed into the network and processed through a 7×7 convolutional layer (Conv1) to output 64 feature maps, which are then downsampled using a stride of 2 convolution operation. Next, a 3×3 Max Pooling layer is employed to additionally halve the dimensions of the feature map.
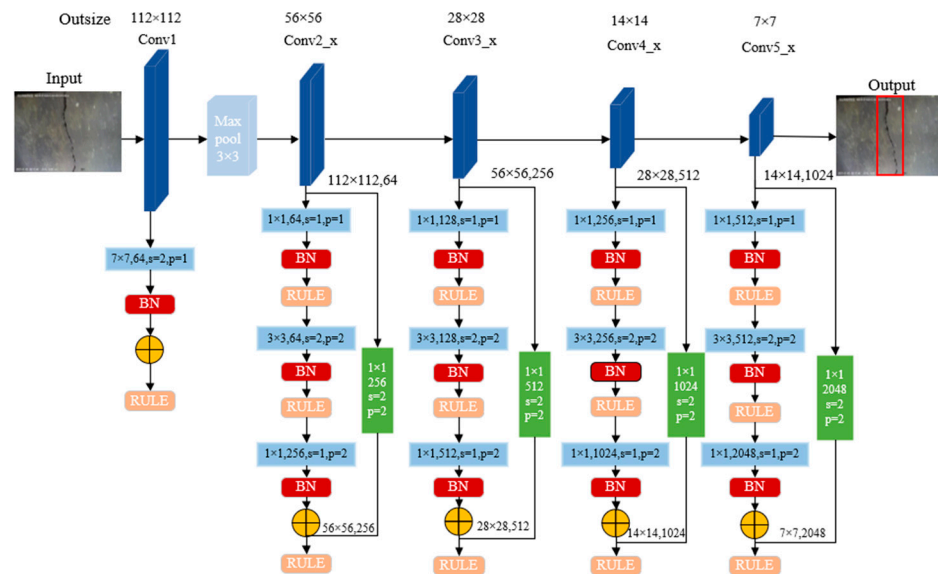

Figure 2. Resnet101 framework.

The core part of the network consists of four main convolution modules, namely Conv2_x, Conv3_x, Conv4_x, and Conv5_x. Each module adopts a "bottleneck" residual structure (Bottleneck Block) based on three convolutional layers, including 1×1, 3×3, and 1×1 convolution operations. The 1×1 convolution is commonly used for channel dimension reduction and restoration, while the 3×3 convolution is used for feature extraction. In each bottleneck block, residual connections facilitate the direct addition of input to the output, forming a residual path that effectively mitigates the gradient vanishing problem. The connections between each module are implemented using convolutional layers with a stride of 2 to downsample the feature map, such as the first 1×1 convolutional layer in modules Conv2_x and Conv3_x, which has a stride of 2 and gradually expands

the number of channels from 64 to 2048. In the last module, Conv5_x, the output feature map has a spatial resolution of 7×7, and the number of channels is expanded to 2048. Finally, the feature maps of each channel are compressed into a single value through Global Average Pooling and subsequently fed into the fully connected layer for output classification prediction.

Several standard functional blocks are integrated into the convolutional modules illustrated in Figure 2 to improve training efficiency and enhance the stability of the model. In particular, Batch Normalization (BN) layers are employed to normalize layer inputs, effectively accelerating network convergence and mitigating internal covariate shifts during training. Additionally, Rectified Linear Unit (ReLU) functions are incorporated as activation units to introduce non-linear transformations, thereby preventing gradient vanishing. The coordinated operation of these components significantly improves the model's representation capabilities and overall robustness during the training process.

The entire network demonstrates high robustness and accuracy in extracting and processing features of concrete cracks, especially in capturing complex crack morphologies. ResNet101, as the backbone network of transFissNet, ensures high precision and reliability in the segmentation task.

3.2. transFissNet Architecture

The proposed transFissNet is a deep learning model specifically tailored for the task of concrete crack segmentation. Figure 3 shows the overall architecture of our proposed transFissNet network, which integrates the multi-scale feature extraction capabilities of convolutional neural networks with the global context modeling power of Transformer modules. To enhance performance in complex scenarios, the architecture incorporates a Residual Attention (RA) module that highlights crack-relevant features while suppressing background noise through attention mechanisms. In the decoder, a Progressive Detail Decoder (PDD) progressively restores spatial detail via multi-level fusion and upsampling, improving boundary delineation. These modules are essential for handling low-contrast and irregular crack patterns.
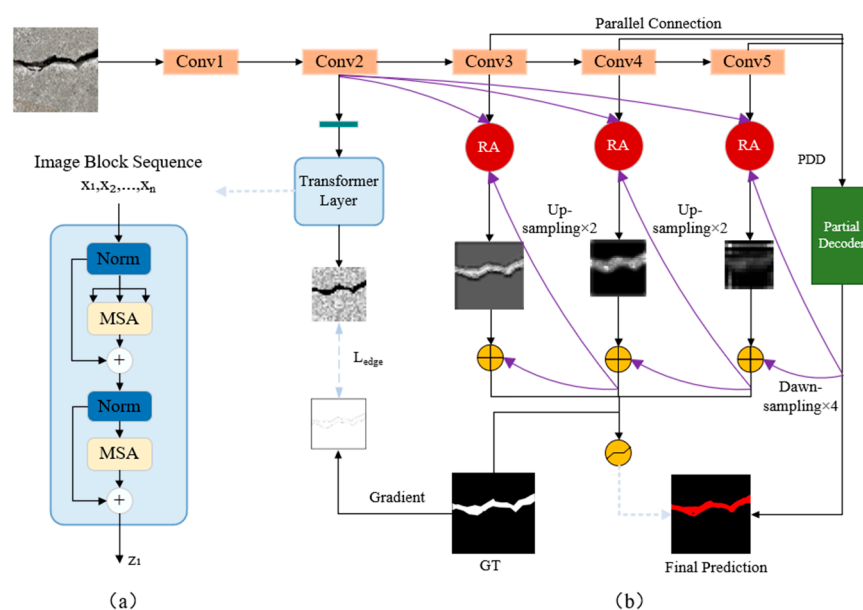

Figure 3. Segmentation model (**a**) schematic of the Transformer layer (**b**) architecture of the transFissNet.

The Transformer module consists of two layers of multi-head self-attention and feed-forward submodules, each followed by layer normalization and residual connections. It is embedded after the encoder to model global spatial dependencies and refine feature

representations before decoding. The input to the Transformer is the flattened and position-encoded feature map obtained from the last convolutional block of the encoder.

TransFissNet employs a customized multi-scale feature extraction strategy based on a ResNet101 backbone, which enables the network to effectively capture crack features at different scales through convolutional and downsampling layers. This design addresses the morphological diversity of cracks and improves the network's adaptability to various real-world conditions. To further enhance long-range dependency modeling, a Transformer module is embedded in the network, using a self-attention mechanism to simulate complex spatial relationships among crack regions. This improves the model's capability to identify ambiguous or weakly defined crack boundaries. The segmentation output of transFissNet serves as the input to the subsequent defect assessment and repair decision modules, enabling intelligent crack-level analysis and downstream repair planning.

This study also introduces specific design considerations for both the encoder and decoder. The encoder of transFissNet leverages the pre-trained deep convolutional neural network, ResNet101. After deep optimization, the model can effectively extract multi-scale features of the input image. Through a sequence of convolutional layers, batch normalization layers, and activation function operations, the encoder progressively extracts high-level features from the crack image. In this design of multi-scale feature extraction, the model can better cope with complex crack shapes and backgrounds while still effectively capturing the details and global structure of cracks. The decoder adopts a multi-level lateral output structure. Following certain column upsampling operations, the feature maps at various levels in the encoder are integrated with the current upsampling outcome via skip connections. This feature fusion not only preserves high-resolution spatial information but also enhances the ability to locate crack boundaries accurately. Ultimately, the segmentation image generated by the decoder can clearly depict the details and morphology of cracks, effectively addressing complex crack segmentation tasks.

4. Experiments and Results

4.1. Datasets

4.1.1. Concrete Crack Dataset

The image dataset employed for segmentation in this article comprises 437 concrete cracks that were captured using a high-resolution camera. The cracks were collected from multiple real-world concrete structures, including pavements, walls, and sections of exposed pipelines, to ensure a wide range of morphological and environmental characteristics. Taking into account the variety of crack morphologies and external factors, such as lighting conditions in concrete crack segmentation work, we focus on building a more representative and diverse crack image dataset during the data collection process. As illustrated in Figure 4, the images we captured encompass a range of crack formations and lighting conditions, which diminishes the model's reliance on a single scenario and boosts its generalization capability in intricate surroundings.

Additionally, to boost the diversity of the dataset and minimize model bias because of inadequate samples, we implemented data augmentation methods, including rotation, scaling, translation, and flipping on the gathered dataset, expanding it to 2185 images to improve its size and diversity. The training and testing sets are divided in a 7:3. Figure 5: Process of crack image annotation shows the process of annotating the collected dataset.
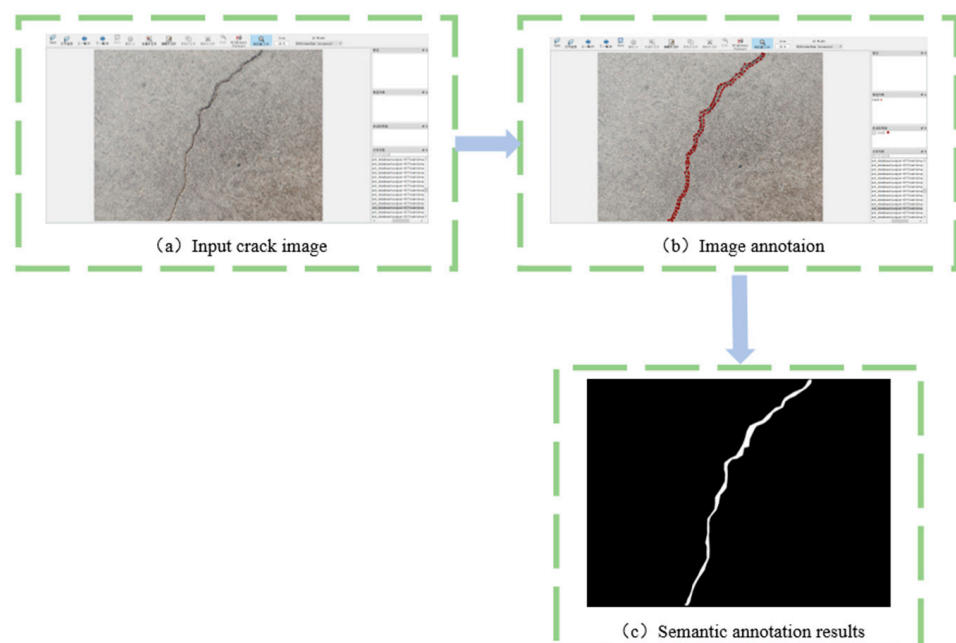
Figure 4. Samples from the dataset.

Figure 5. Process of crack image annotation.

4.1.2. Crack500 Dataset

Yang et al. [48] gathered 500 images of cracks in asphalt pavement at Tianpu University (See Figure 6), each with a resolution of around 2000×1500 pixels, and meticulously labeled them on a pixel-by-pixel basis. Owing to the computer's limited GPU memory, we segmented each original image into five smaller segments and normalized the resolution of the segmented images to 256×256 pixels. Subsequently, the dataset was split into a 70% training set and a 30% testing set, containing 1648 and 707 images, respectively. The images in this dataset cover not only various crack shapes but also complex environmental conditions such as shadows and uneven lighting, enhancing the model's capacity to generalize in crack segmentation. Although splitting images into patches may reduce global contextual information, this limitation is mitigated by the encoder's multi-scale feature extraction and the Transformer module's global attention mechanism, which together enable the model to capture long-range dependencies even from locally constrained inputs.
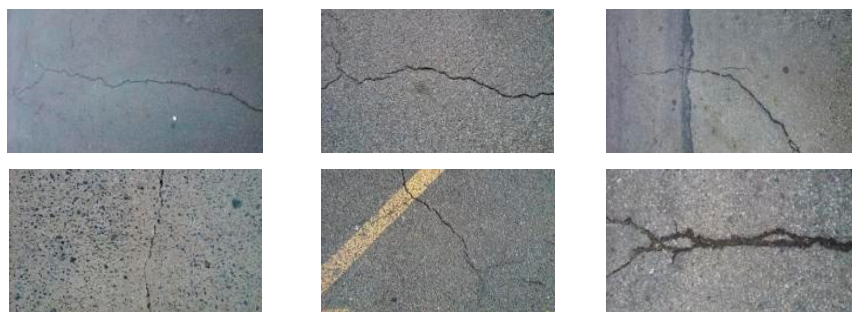
Figure 6. Samples of the Crack500 dataset.

4.1.3. CrackForest Dataset

As illustrated in Figure 7, the CrackForest dataset, which is dedicated to the research on the detection and segmentation of concrete cracks, has been published by Tianjin University. The dataset consists of 118 labeled crack images, each with a resolution of 480×320 pixels, aimed at preventing overfitting and improving the model's generalization capability; this dataset enhances the original crack images through methods such as blurring, brightness enhancement or reduction, 180 degree rotation, and horizontal mirroring, creating 590 images depicting cracks along with their respective annotated versions. The total number of samples for data collection was further increased to 708, with 420 images being arbitrarily picked for the training set, 144 for the validation set, and 144 for the testing set.

Figure 7. Samples of the CFD dataset.

4.2. Experimental Setup

All experiments were conducted on a workstation equipped with an NVIDIA GeForce RTX 2080 SUPER GPU(NVIDIA Corporation, Santa Clara, CA, USA), using the PyCharm version 2023.2.2 environment for model development and training. Prior to training, all input images were preprocessed to ensure quality and consistency. This preprocessing step included resizing all images to a uniform resolution of 204×204 pixels, removing blurry or low-quality samples, and adjusting brightness and contrast to normalize the lighting conditions across the dataset. These procedures ensured that these input data maintained consistent quality standards and eliminated potential interference caused by poor image conditions. To mitigate overfitting and improve data diversity, a series of data augmentation techniques were applied to the collected dataset, including brightness enhancement and attenuation, saturation adjustment, and horizontal flipping. As a result, the dataset was expanded to 2185 images, from which 1355 high-quality images were selected for training and testing, with 948 images used for training and 407 images used for testing, maintaining a 7:3 ratio. The training process used a batch size of 16 and a learning rate of 0.0001. The model was trained for 100 epochs. The loss function is a combination of binary cross-entropy (BCE) and intersection-over-union (IoU) losses. The BCE term addresses pixel-level accuracy, especially by focusing on the crack pixels in the presence of class imbalance. The IoU term improves segmentation consistency by optimizing the

overlap between the predicted and true crack regions. Together, these losses ensure both fine-grained accuracy and robust region-level segmentation.

4.3. Optimizer

Within this research study, the model was trained utilizing the Adam (Adaptive Moment Estimation) optimization algorithm. The Adam optimization algorithm integrates the benefits of the momentum approach and adaptive learning rate optimization, updating model parameters by calculating the gradients' first-order moment, which is the mean, and the second-order moment, which is the variance. Compared with traditional stochastic gradient descent (SGD) algorithms, Adam exhibits faster convergence speed and better robustness when dealing with high-dimensional data with sparse gradients. The excellent performance of the Adam optimizer makes it one of the preferred optimization algorithms in various deep learning tasks, particularly ideal for addressing extensive parameter tuning challenges within this research.

4.4. Evaluating Metrics

Our network functions as a semantic segmentation model capable of distinguishing concrete cracks from the surrounding background. After inputting the image, it can predict the output of the binary mask image through pre-trained knowledge. To assess the model's performance in a more impartial and precise manner, we used four widely used evaluation metrics in this type of task, including Accuracy, Recall, Mean Dice Coefficient, and S-measure. We categorize the pixels representing cracks in each image as positive samples, while those depicting the background are considered negative samples.

The definitions of the four assessment criteria are as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$meanDic = \frac{1}{N} \sum_{i=1}^{N} \frac{2 \times |A_i \cap B_i|}{|A_i| + |B_i|} \tag{3}$$

$$S\text{-}measure = \alpha \times S_R + (1 - \alpha) \times S_O \tag{4}$$

TP denotes the count of true positives, TN signifies the count of true negatives, FP indicates the count of false positives, FN stands for the count of false negatives, A_i represents the predicted result of the i-th sample, B_i denotes the actual label for the i-th instance, α represents the weight factor, S_R represents the accuracy of the segmentation result within the entire region, and S_O represents the similarity in shape and structure between the segmentation result boundary and the true label boundary.

4.5. Comparative Experiments

To make an unbiased comparison of the performance across various networks, we train all networks under the same operating environment and parameter settings until convergence and optimal performance are achieved. Table 1 presents a performance comparison of various leading deep learning networks (such as DeepLabv3, PSPNet, UNet, transUNet, and FCN) in concrete crack segmentation tasks, with evaluation metrics including Accuracy, Recall, meanDice, and S-measure. As observed in the table, the transFissNet introduced in this study surpasses other networks in the majority of evaluation metrics, demonstrating excellent segmentation performance. Specifically, transFissNet has achieved significant advantages in accuracy, reaching 0.968, which is significantly

better than FCN's 0.924, UNet's 0.876, and PSPNet's 0.796, indicating that transFissNet can provide higher classification accuracy in crack detection. In addition, in terms of the key indicators meanDice and S-measure, which measure segmentation accuracy and consistency, transFissNet achieved the highest values of 0.696 and 0.802 among all models, further demonstrating its excellent ability to capture complex crack shapes and maintain edge consistency.

Table 1. Comparison of results obtained from various network models.

The Utilized Network	Accuracy	Recall	meanDic	S-Measure
FCN	0.924	0.915	0.671	0.781
UNet	0.876	0.841	0.653	0.779
transUNet	0.903	0.859	0.664	0.784
DeepLabV3	0.757	0.702	0.592	0.776
PSPNet	0.796	0.728	0.605	0.751
transFissNet	0.968	0.904	0.696	0.802

From Figure 8, the trends in accuracy and loss for each network throughout the training process are observable. Figure 8a illustrates the fluctuation in accuracy for various networks across different iterations throughout the training phase. As depicted by the graph, transFissNet demonstrates a swift enhancement in accuracy during the initial phases of training, gradually stabilizing after several iterations and reaching the highest accuracy near 40 epochs. In contrast, other networks such as PSPNet and DeepLabV3 have a slower increase in accuracy during the training process, especially in more complex crack segmentation tasks, where they have not achieved the same level of accuracy as transFissNet. Figure 8b illustrates the curves of loss reduction for each network. It is observable that transFissNet exhibits a swift decline in loss values, attaining a lower plateau during the initial phases, indicating that its convergence speed is faster and its stability is better. The loss reduction in other networks (such as DeepLabV3 and PSPNet) is relatively small, and their convergence performance in the later stages of training is relatively weak, reflecting their poor performance in complex crack detection tasks.
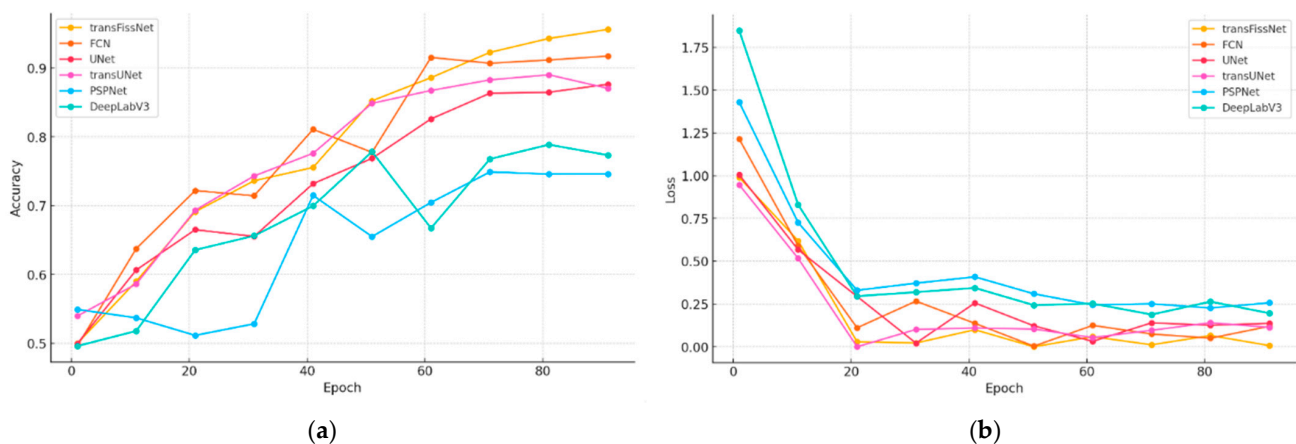

(**a**) (**b**)

Figure 8. The performance of the network in the training process: (**a**) Accuracy; (**b**) Loss.

Although FCN has a slight advantage in Recall (0.915), transFissNet still leads in overall performance. Especially in terms of Accuracy and meanDice, the results demonstrate that transFissNet achieves a better balance in identifying and segmenting crack areas. While FCN performs well on average, it tends to struggle in scenarios involving blurred boundaries and complex crack structures. This highlights the advantage of transFissNet, which can model global context and adapt to irregular morphologies through its Transformer-enhanced structure. In addition, DeepLabV3 and PSPNet are relatively weak in handling

complex scenes, especially in terms of performance on meanDice and S-measure, which are 0.592 and 0.605, respectively, indicating their limitations in dealing with small-scale cracks and complex backgrounds. Overall, transFissNet has demonstrated stronger capabilities in concrete crack detection and segmentation tasks and can better handle the complex morphology and background interference of cracks, significantly outperforming other networks.

Figure 9 further compares the accuracy and recall of different networks in the final segmentation task. It can be seen that transFissNet performs well in both accuracy and recall, especially in accuracy, where it maintains a leading advantage, indicating that it not only has high classification accuracy in crack detection but also achieves a relatively balanced performance without sacrificing recall. In comparison, DeepLabV3 and PSPNet perform relatively poorly on these two metrics, particularly in Recall, where they fail to capture more crack areas effectively.
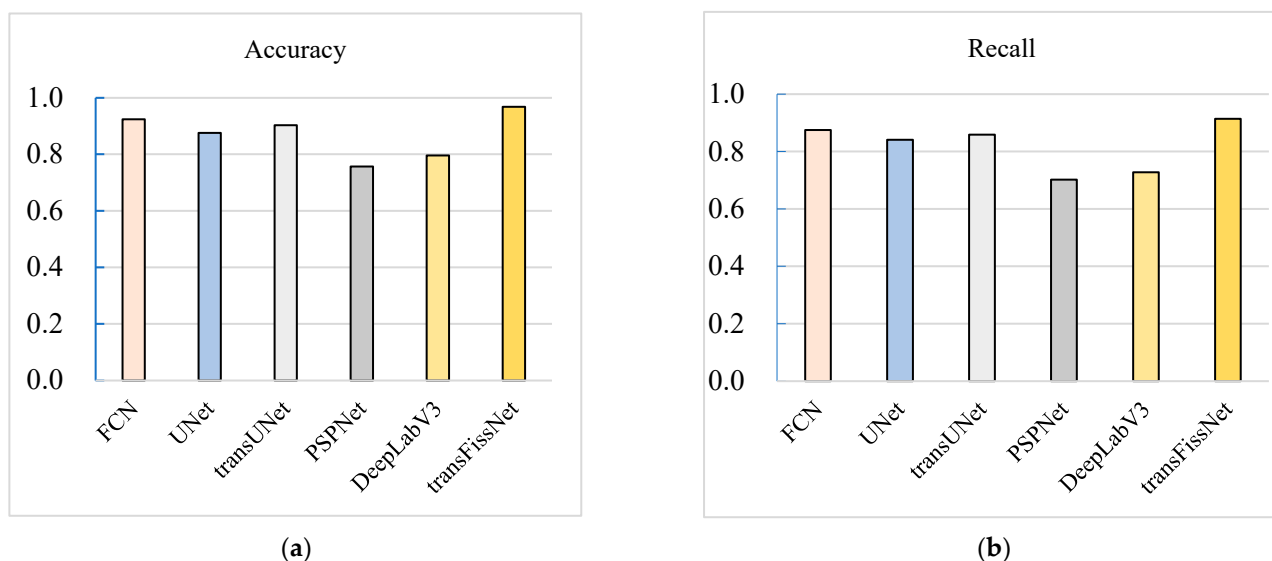

Figure 9. The crack segmentation performance of different networks (**a**) Accuracy; (**b**) Recall.

Figure 10 illustrates the visual outcomes of various deep learning models when segmenting concrete cracks, further illustrating the actual performance of each network in handling crack morphologies and background interference. The figure includes Raw Images, GT, and segmentation outputs from six networks (DeepLabV3, PSPNet, UNet, TransUNet, FCN, and transFissNet). It can be observed that transFissNet excels at capturing crack details and edge processing, especially in cases where cracks are complex and have irregular shapes, with segmentation results closely matching the GT. In contrast, the segmentation results of DeepLabV3 and PSPNet are relatively blurry in images with complex crack shapes, with significant loss of detail. UNet and TransUNet perform well in simple scenarios but experience a decline in accuracy in images with highly curved cracks. FCN shows poor performance in handling crack edges, with rough segmentation results and a lack of detailed boundary recognition. The figure validates the advantages of transFissNet in crack segmentation tasks, effectively addressing the complexity of crack morphologies and providing more accurate segmentation results.
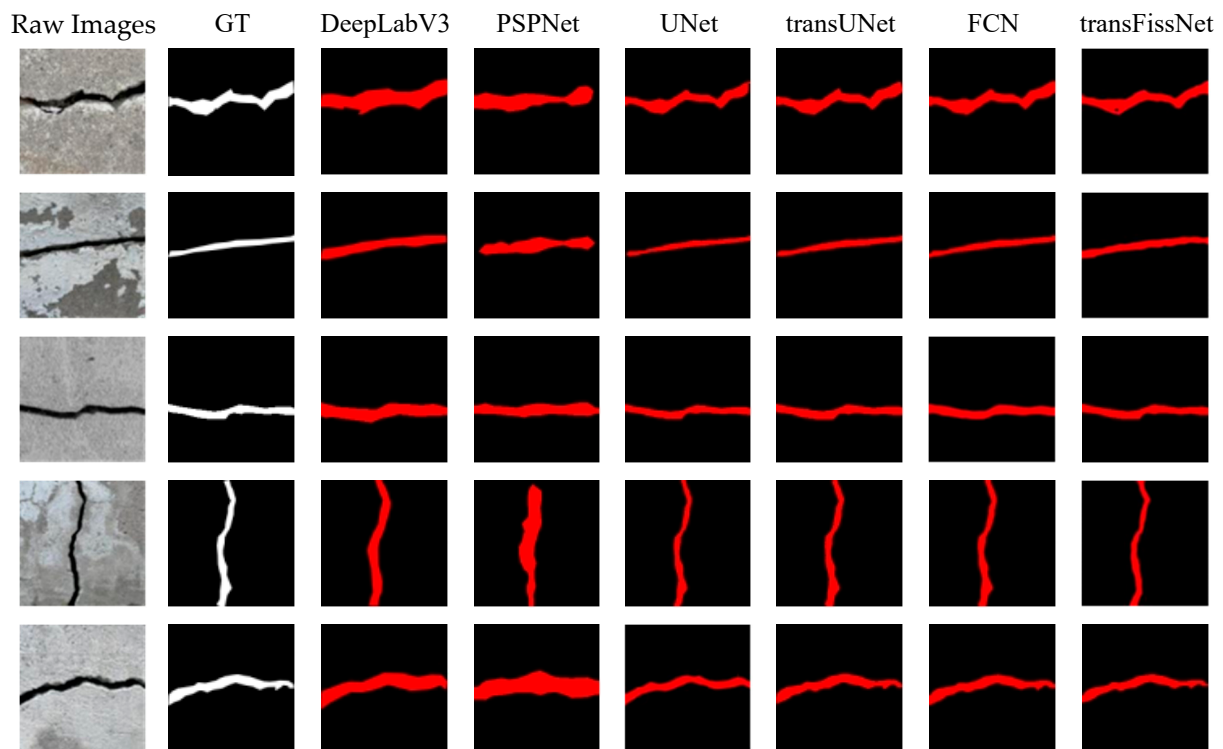
Figure 10. The crack segmentation results of different networks on our dataset.

4.6. Crack Identification System

In this study, we proposed an urban pipeline crack defect identification and repair system based on image processing and intelligent analysis technology, aiming to enhance the intelligent management level of pipeline and concrete crack defects (See Figure 11). The system consists of four core modules: electronic reporting, defect identification, defect classification, and intelligent repair plan generation, with functions including data presentation, image analysis, decision support, and report output.
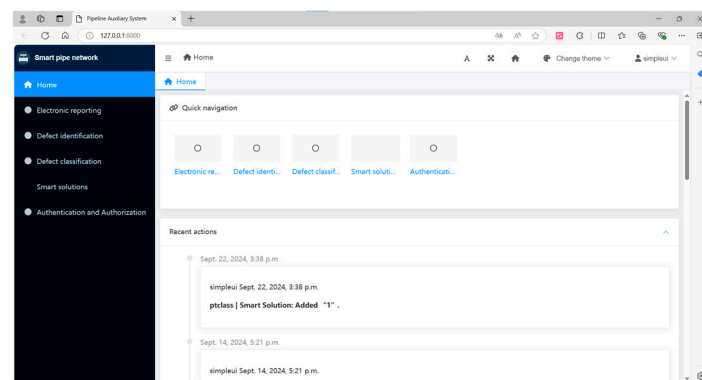
Figure 11. System homepage.

In response to the challenges associated with station-based total stations—such as difficulties in vehicle navigation, high labor and time costs, and structural incompatibility—we propose an intelligent engineering vehicle system that integrates total stations, laser scanners, and dual-spectrum gimbal sensors. This system is designed to enhance the efficiency and accuracy of quality perception by leveraging the respective strengths of these sensing devices. As illustrated in Figure 12a, a vehicle-mounted system incorporating a modular total station and laser scanner was developed.
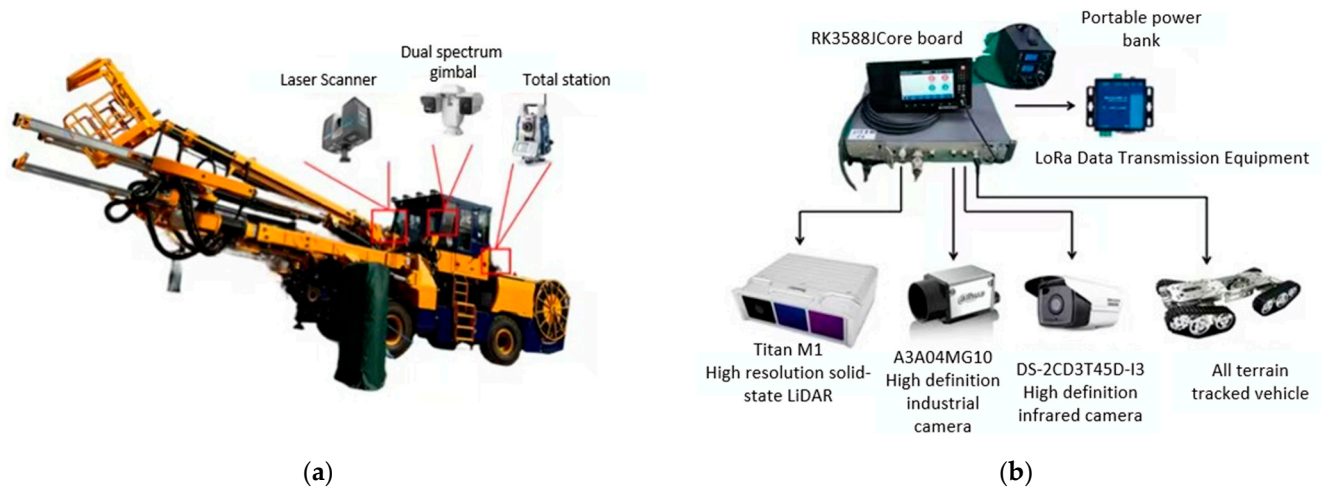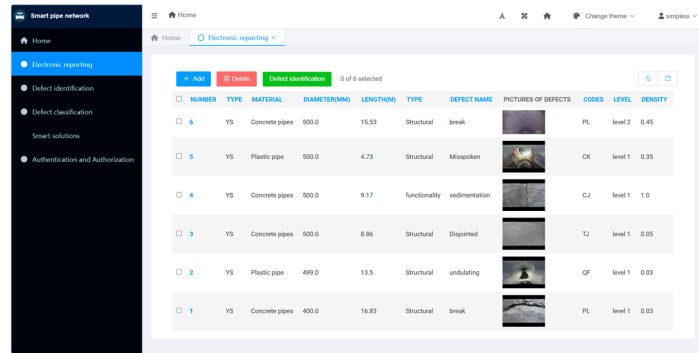
Figure 12. Dataset collection robot (**a**) Data acquisition equipment (**b**) Driven equipment.

To achieve precise diagnostics of concrete cracks, we created an intelligent diagnostic instrument that integrates sensing, data fusion, and diagnostic functions. This device utilizes 3D LiDAR, high-definition industrial cameras, and infrared sensors, all mounted on an all-terrain tracked vehicle to improve maneuverability in complex environments. The intelligent diagnostic instrument is composed of five integrated components, as shown in Figure 12b. The tracked vehicle is outfitted with 3D LiDAR, industrial cameras, infrared sensors, and a central processing unit. It also features a belt-driven transmission system, shock absorption, a power supply, an automatic lifting mechanism, and a rotational control unit. The control system manages vehicle movement and adjusts the lighting compensation system. System debugging was carried out to ensure effective communication and data transmission between components, confirming stability and accuracy. The algorithms were further optimized to adapt to various tunnel environments and working conditions. Additionally, point cloud data obtained by LiDAR were processed using a least squares fitting algorithm, enabling precise measurement and modeling of concrete structures.

As illustrated in Figure 13, the electronic reporting module manages and displays original detection data of crack defects in a standardized table format, supporting multi-dimensional data filtering and querying, which facilitates users to locate the target defect records quickly. The defect identification module introduces the transFissNet model, combined with image preprocessing and data enhancement techniques, to achieve automatic recognition and boundary positioning of defect areas in detection images, enhancing the accuracy of crack defect identification in complex environments(See Figure 14). The identification results are presented in a visual manner, supporting manual verification and in-depth analysis. The defect classification module classifies crack defects into fine granularity based on the morphological characteristics of the identified areas and combines multi-dimensional parameters such as pipeline materials, defect shapes, and distribution characteristics to provide a basis for generating accurate repair plans.

The repair plan generation module integrates the results of defect segmentation and classification, as well as information such as defect location, pipeline network materials, service life, and pipeline operating conditions, to automatically generate repair plans using rule-based reasoning and data-driven models. The plan includes repair measures (such as local reinforcement or complete replacement), required materials and equipment, construction period estimation, and cost assessment, and can be optimized according to different constraints.

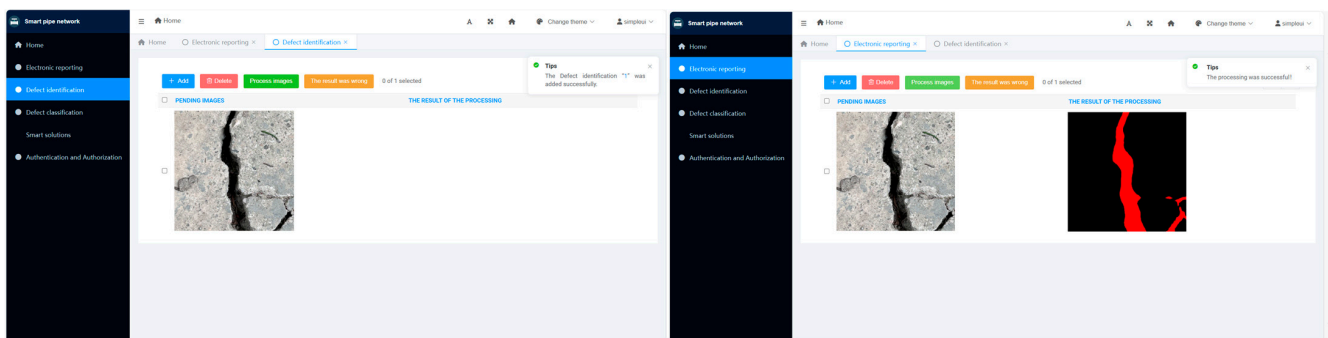**Figure 13.** Electronic report module.

Figure 14. Defect recognition module.

By integrating the cost–benefit ratio calculations and plan evaluation models of various repair strategies, the system can iteratively generate the optimal repair plan. Finally, the system can automatically generate standardized repair reports, supporting export and sharing, providing data support and a decision-making basis for the entire lifecycle management of the pipeline network. The system helps to improve the automation level of pipeline defect identification and the intelligence level of repair decision-making, improves the efficiency and precision of management, and offers vital technical backing for the enduring secure functioning of urban pipeline systems.

5. Conclusions

In this paper, we proposed an intelligent concrete crack segmentation network trans-FissNet based on ResNet101 and Transformer modules. By integrating the multi-scale feature extraction capability of convolutional neural networks with the global information modeling of Transformers, efficient crack detection and segmentation are achieved. The custom encoder–decoder structure and self-attention mechanism in the model significantly enhance the ability to capture long-range dependencies of cracks under complex backgrounds, while the multi-level feature fusion strategy strengthens the precise recognition of crack edge details. Experimental results show that transFissNet outperforms existing mainstream algorithms on multi-scenario crack image datasets, with outstanding performance in accuracy, recall rate, mean Dice coefficient, and S-measure, especially in situations with diverse crack morphologies and complex background interference. The results prove the adaptability and robustness of transFissNet in crack segmentation, greatly improving detection accuracy and practicality.

Future optimization directions include: (1) Further reducing the computational complexity of the model to improve its deployment efficiency in resource-constrained environments. (2) Improving inference speed while maintaining accuracy. (3) Expanding the diversity of the training dataset, especially by adding crack data simulated under

complex working conditions to enhance generalization ability. (4) Conducting in-depth research on crack classification, performing detailed analysis on different types of cracks that pose structural hazards, and providing more precise support for structural health monitoring. (5) Exploring the integration of crack depth information, whether obtained through manual measurement or 3D sensing, to enable more comprehensive structural assessments. Through further optimization of the model architecture and expansion of application scenarios, transFissNet is expected to become an important tool for intelligent structure inspection and maintenance in the future, providing a more efficient and accurate solution for crack monitoring in complex engineering environments.

Author Contributions: T.Y.: Methodology, Investigation, Data Curation. M.H.: Supervision, Funding Acquisition. Y.W.: Formal Analysis, Validation, Writing—review and editing. J.W.: Validation, Visualization. L.Z.: Data Curation. J.Z.: Supervision, Writing—review and editing, Resources. All authors have read and agreed to the published version of the manuscript.

Data Availability Statement: The datasets used in this study are publicly available. The CrackForest dataset can be accessed at https://github.com/cuilimeng/CrackForest (accessed on 16 April 2025), and the Crack500 dataset is available at https://data.mendeley.com/datasets/wddt4gbttd (accessed on 16 April 2025). Both datasets are licensed under the Creative Commons Attribution 4.0 International License (CC BY 4.0). Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Golewski, G.L. The phenomenon of cracking in cement concretes and reinforced concrete structures: The mechanism of cracks formation, causes of their initiation, types and places of occurrence, and methods of detection—A review. *Buildings* **2023**, *13*, 765. [CrossRef]
2. Elmenshawy, Y.; Elmahdy, M.A.; Moawad, M.; Elshami, A.A.; Ahmad, S.S.; Nagai, K. Investigating the bacterial sustainable self-healing capabilities of cracks in structural concrete at different temperatures. *Case Stud. Constr. Mater.* **2024**, *20*, e03188. [CrossRef]
3. Hong, H.F.; Xi, W. Causes and Control Measures of Mass Concrete Crack of High-Rise Building Basement Foundation Slab. *Adv. Mater. Res.* **2010**, *163–167*, 1609–1613.
4. Lu, W. Causes of Concrete Cracks in Civil Construction and Prevention Measures. *Soc. Responsib. Acad. Progress.* **2020**, *2*, 25–26.
5. Xu, G.; Han, X.; Zhang, Y.; Wu, C. Dam Crack Image Detection Model on Feature Enhancement and Attention Mechanism. *Water* **2022**, *15*, 64. [CrossRef]
6. Ge, Y.; Liu, J.; Zhang, X.; Tang, H.; Xia, X. Automated Detection and Characterization of Cracks on Concrete Using Laser Scanning. *J. Infrastruct. Syst.* **2023**, *29*, 04023005. [CrossRef]
7. Chun, Z.; Jian, Y.; Ruoqing, W. An enhanced crack segmentation method using implicit classification and inference rules in steel bridge. *Vis. Comput.* **2024**, *40*, 4001–4021.
8. Liang, F.; Li, Q.; Chai, X.; Wang, W. A two-stage unsupervised crack image segmentation method. *J. Beijing Jiaotong Univ.* **2023**, *47*, 122–128.
9. Chen, L.; Zhou, L.; Li, L.; Luo, M. CrackDiffusion: Crack inpainting with denoising diffusion models and crack segmentation perceptual score. *Smart Mater. Struct.* **2023**, *32*, 054001. [CrossRef]
10. Zhou, S.; Pan, Y.; Huang, X.; Yang, D.; Ding, Y.; Duan, R. Crack Texture Feature Identification of Fiber Reinforced Concrete Based on Deep Learning. *Materials* **2022**, *15*, 3940. [CrossRef]
11. Ding, Y.; Zhou, S.-X.; Yuan, H.-Q.; Pan, Y.; Dong, J.-L.; Wang, Z.-P.; Yang, T.-L.; She, A.-M. Crack Identification Method of Steel Fiber Reinforced Concrete Based on Deep Learning: A Comparative Study and Shared Crack Database. *Adv. Mater. Sci. Eng.* **2021**, *2021*, 9934250. [CrossRef]
12. Zhu, Y.; Tang, H. Automatic Damage Detection and Diagnosis for Hydraulic Structures Using Drones and Artificial Intelligence Techniques. *Remote Sens.* **2023**, *15*, 615. [CrossRef]
13. Deng, L.; Chu, H.-H.; Shi, P.; Wang, W.; Kong, X. Region-Based CNN Method with Deformable Modules for Visually Classifying Concrete Cracks. *Appl. Sci.* **2020**, *10*, 2528. [CrossRef]

14. Dung, V.C.; Anh, D.L. Autonomous concrete crack detection using deep fully convolutional neural network. *Autom. Constr.* **2019**, *99*, 52–58. [CrossRef]

15. Ham, S.; Bae, S.; Lee, I.; Lee, G.P.; Kim, D. An evaluation methodology for cement concrete lining crack segmentation deep learning model. *J. Korean Tunn. Undergr. Space Assoc.* **2022**, *24*, 513–524.

16. Zhao, X.; Ni, Y. Object Component Segmentation Network Based on DeepLab. *Pattern Recognit. Artif. Intell.* **2020**, *33*, 211–220.

17. Lei, Q.; Zhong, J.; Wang, C.; Li, X. Integrating Crack Causal Augmentation Framework and Dynamic Binary Threshold for imbalanced crack instance segmentation. *Expert Syst. Appl.* **2024**, *240*, 122552. [CrossRef]

18. Liu, H.; Yang, J.; Miao, X.; Mertz, C.; Kong, H. Crackformer network for pavement crack segmentation. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 9240–9252. [CrossRef]

19. Wu, Y.; Li, S.; Zhang, J.; Li, Y.; Li, Y.; Zhang, Y. Dual attention transformer network for pixel-level concrete crack segmentation considering camera placement. *Autom. Constr.* **2024**, *157*, 105166. [CrossRef]

20. Zhou, J.; Zhao, G.; Li, Y. Vison Transformer-Based Automatic Crack Detection on Dam Surface. *Water* **2024**, *16*, 1348. [CrossRef]

21. Yuan, M.; Huang, H.; Zhou, C. Research progress of image semantic segmentation based on fully supervised learning. *Comput. Eng. Appl.* **2021**, *57*, 43–54.

22. Shi, Z.; Jin, N.; Chen, D.; Ai, D. A comparison study of semantic segmentation networks for crack detection in construction materials. *Constr. Build. Mater.* **2024**, *414*, 134950. [CrossRef]

23. Kim, B.; Kim, G.; Jin, S.; Cho, S. A comparative study on performance of deep learning models for vision-based concrete crack detection according to model types. *J. Korean Soc. Saf.* **2019**, *34*, 50–57.

24. Stone, J.A.; Maynard, I.W.; North, J.S.; Panchuk, D.; Davids, K. (De)synchronization of advanced visual information and ball flight characteristics constrains emergent information-movement couplings during one-handed catching. *Exp. Brain Res.* **2015**, *233*, 449–458. [CrossRef]

25. Zhang, J.; Xiang, K.; Wang, J.; Liu, J.; Kang, M.; Pan, Z. Trans-Inf-Net: COVID-19 Lung Infection Segmentation Based on Transformer. In Proceedings of the 2022 8th International Conference on Virtual Reality (ICVR), Nanjing, China, 26–28 May 2022.

26. Liang, D.; Li, Y.; Zhang, S. Identification of cracks in concrete bridges through fusing improved ResNet-14 and RS-Unet models. *J. Beijing Jiaotong Univ.* **2023**, *47*, 10–18.

27. Xu, G.; Yue, Q.; Liu, X.; Chen, H. Investigation on the effect of data quality and quantity of concrete cracks on the perfor-mance of deep learning-based image segmentation. *Expert Syst. Appl.* **2024**, *237*, 121686. [CrossRef]

28. Meng, X. Concrete Crack Detection Algorithm Based on Deep Residual Neural Networks. *Sci. Program.* **2021**, *2021*, 3137083. [CrossRef]

29. Teng, S.; Liu, Z.; Chen, G.; Cheng, L. Concrete crack detection based on well-known feature extractor model and the YOLO_v2 network. *Appl. Sci.* **2021**, *11*, 813. [CrossRef]

30. Ye, G.; Li, S.; Zhou, M.; Mao, Y.; Qu, J.; Shi, T.; Jin, Q. Pavement crack instance segmentation using YOLOv7-WMF with connected feature fusion. *Autom. Constr.* **2024**, *160*, 105331. [CrossRef]

31. Wan, C.; Xiong, X.; Wen, B.; Gao, S.; Fang, D.; Yang, C.; Xue, S. Crack detection for concrete bridges with imaged based deep learning. *Sci. Prog.* **2022**, *105*, 00368504221128487. [CrossRef]

32. Kim, B.; Soo, C. Image-based concrete crack assessment using mask and region-based convolutional neural network. *Struct. Control Health Monit.* **2019**, *26*, e2381. [CrossRef]

33. Xu, G.; Yue, Q.; Liu, X. Deep learning algorithm for real-time automatic crack detection, segmentation, qualification. *Eng. Appl. Artif. Intell.* **2023**, *126*, 107085. [CrossRef]

34. Liu, Y.; Zhou, T.; Xu, J.; Hong, Y.; Pu, Q.; Wen, X. Rotating Target Detection Method of Concrete Bridge Crack Based on YOLO v5. *Appl. Sci.* **2023**, *13*, 11118. [CrossRef]

35. Liu, F.; Liu, J.; Wang, L.; Al-Qadi, I.L. Multiple-type distress detection in asphalt concrete pavement using infrared thermography and deep learning. *Autom. Constr.* **2024**, *161*, 105355. [CrossRef]

36. Guo, L.; Li, R.; Jiang, B. A cascade broad neural network for concrete structural crack damage automated classification. *IEEE Trans. Ind. Inform.* **2020**, *17*, 2737–2742. [CrossRef]

37. Hacıefendioğlu, K.; Başağa, H.B. Concrete road crack detection using deep learning-based faster R-CNN method. *Iran. J. Sci. Technol. Trans. Civ. Eng.* **2022**, *46*, 1621–1633.

38. Shi, J.; Liu, C.; Wang, D.; Liu, Z.; Liu, G.; Chun, Q.; Zhang, Y. Numerical simulation of effective diffusivity in concrete with random microcracks. *J. Build. Eng.* **2023**, *63*, 105501. [CrossRef]

39. Tang, J.; Chen, C.; Huang, Z.; Zhang, X.; Li, W.; Huang, M.; Deng, L. Crack Unet: Crack Recognition Algorithm Based on Three-Dimensional Ground Penetrating Radar Images. *Sensors* **2022**, *22*, 9366. [CrossRef]

40. Su, C.; Wang, W. Concrete Cracks Detection Using Convolutional Neural Network Based on Transfer Learning. *Math. Probl. Eng.* **2020**, *2020*, 7240129. [CrossRef]

41. Wang, Q.; Chen, B. A novel transfer learning model for the real-time concrete crack detection. *Knowl.-Based Syst.* **2024**, *301*, 112313.

42. Fang, J.; Yang, C.; Shi, Y.; Wang, N.; Zhao, Y. External attention based TransUNet and label expansion strategy for crack detection. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 19054–19063. [CrossRef]

43. Philip, R.E.; Andrushia, A.D.; Nammalvar, A.; Gurupatham, B.G.A.; Roy, K. A Comparative Study on Crack Detection in Concrete Walls Using Transfer Learning Techniques. *J. Compos. Sci.* **2023**, *7*, 169. [CrossRef]

44. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 19–25 June 2021; pp. 1657–1667.

45. Xu, Y.; Li, D.; Xie, Q.; Wu, Q.; Wang, J. Automatic defect detection and segmentation of tunnel surface using modified Mask R-CNN. *Measurement* **2021**, *178*, 109316. [CrossRef]

46. Ranyal, E.; Sadhu, A.; Jain, K. Enhancing pavement health assessment: An attention-based approach for accurate crack detection, measurement, and mapping. *Expert Syst. Appl.* **2024**, *247*, 123314. [CrossRef]

47. Yao, H.; Wu, Y.; Liu, S.; Liu, Y.; Xie, H. A pavement crack synthesis method based on conditional generative adversarial networks. *Math. Biosci. Eng.* **2024**, *21*, 903–923. [CrossRef]

48. Yang, X.; Li, H.; Yu, Y.; Luo, X.; Huang, T. Automatic pixel-level crack detection and measurement using fully convolutional network. *Comput.-Aided Civ. Infrastruct. Eng.* **2019**, *34*, 616–634. [CrossRef]