buildings

MDPI

Article

Recognition of Concrete Surface Cracks Based on Improved TransUNet

Xuwei Dong [1], Yang Liu [1] and Jinpeng Dai [2,3,4,*]

[1] Key Laboratory of Opto-Electronic Technology and Intelligent Control, Ministry of Education, Lanzhou Jiaotong University, Lanzhou 730070, China; dxw007@lzjtu.edu.cn (X.D.); 12221950@stu.lzjtu.edu.cn (Y.L.)
[2] National and Provincial Joint Engineering Laboratory of Road & Bridge Disaster Prevention and Control, Lanzhou Jiaotong University, Lanzhou 730070, China
[3] State Key Laboratory of High Preformance in Civil Engineering Materials, Jiangsu Research Institute of Building Science Co., Ltd., Nanjing 210008, China
[4] School of Materials Science and Engineering, Southeast University, Nanjing 211189, China
* Correspondence: daijp@mail.lzjtu.cn

Abstract: Concrete surface crack detection is a critical problem in the health monitoring and maintenance of engineering structures. The existence and development of cracks may lead to the deterioration of structural performance, potentially causing serious safety accidents. However, detecting cracks accurately remains challenging due to various factors such as uneven lighting, noise interference, and complex backgrounds, which often lead to incomplete or false detections. Traditional manual inspection methods are subjective, inefficient, and costly, while existing deep learning-based approaches still have the problem of insufficient precision and completeness. Therefore, this paper proposes a new crack detection model based on an improved TransUNet: AG-TransUNet, an adaptive multi-head self-attention mechanism, and a gated mechanism-based decoding module (GRU-T) is introduced to improve the accuracy and completeness of crack detection. Experimental results show that the AG-TransUNet outperforms the original TransUNet with a 4.05% increase in precision, a 2.59% improvement in F1-score, and a 0.36% enhancement in IoU on the CFD dataset. The AG-TransUNet achieves a 2.21% increase in precision, a 5.63% improvement in F1-score, and a 9.07% enhancement in IoU on the concrete crack dataset. In addition, in order to further quantitatively analyze the crack width, the orthogonal skeleton method is used to calculate the maximum width of a single crack to provide a reference for engineering maintenance. Experiments show that the maximum error between the real values and detection results is about 5%. Therefore, the proposed method better meets the needs of crack detection in practical engineering applications and provides a solution for improving the efficiency of crack detection.

Keywords: concrete surface cracks; crack detection; semantic segmentation; attention mechanism; crack width

Academic Editor: Marco Di Ludovico

Received: 14 January 2025
Revised: 3 February 2025
Accepted: 8 February 2025
Published: 10 February 2025

Citation: Dong, X.; Liu, Y.; Dai, J. Recognition of Concrete Surface Cracks Based on Improved TransUNet. *Buildings* **2025**, *15*, 541. https://doi.org/10.3390/buildings15040541

1. Introduction

Surface cracks are one of the most common and significant forms of damage in structural health monitoring. Their detection and evaluation are crucial for ensuring the safety and durability of buildings and infrastructure [1]. Over time, structural materials may crack due to various factors such as fatigue, environmental erosion, and load. If these cracks are not identified and treated promptly, they may lead to severe safety hazards and even catastrophic consequences [2]. Traditional manual detection methods are inefficient, subjective, and lack accuracy, making them unable to meet the demands for efficient and

precise crack detection in modern engineering. Accurately detecting the location, shape, and width of cracks is, therefore, vital for ensuring structural safety.

In recent years, crack detection technology has evolved from traditional image processing methods to deep learning-driven automated detection methods. Traditional image processing techniques include edge detection, morphological operation, threshold segmentation, template matching, and frequency domain analysis. While these methods can identify crack edges and shapes to some extent, they often underperform in complex backgrounds. For example, the canny edge detector is a classical algorithm widely used for detecting crack edges by calculating image gradients [3]. Canny et al. [3] proposed an improved edge detection algorithm, which optimized the canny algorithm through filtering and threshold calculation to improve detection accuracy and robustness. Morphological operation is often used to enhance crack features in images. Zhang et al. [4] detected local dim areas containing potential defects from the original images using morphological operation and designed a distance-based shape descriptor to describe numerical features for defect detection. Otsu's thresholding is an automatic threshold selection method that is also widely used in crack detection [5]. Vivekananthan et al. [6] combined the Otsu method with grayscale discrimination to improve the accuracy of crack detection. The template matching-based method [7] detects cracks by comparing them with pre-designed crack templates. Chen et al. [8] adopted a template-matching method based on color feature recognition to optimize automatic target extraction algorithms. Fourier and wavelet transforms are commonly used tools for frequency domain analysis. Gharehbaghi et al. [9] proposed a crack detection method combining wavelet-based feature extraction, feature reduction, and a fast deep learning-based classifier. Crack texture features can be extracted using methods like the gray-level co-occurrence matrix (GLCM). Arya et al. [10] proposed an automatic crack detection method utilizing GLCM. Histogram equalization is commonly used to enhance image contrast, thereby improving crack detection accuracy. Liu et al. [11] proposed a histogram equalization algorithm that fuses the histogram equalization (HE) and the contrast-limited adaptive histogram equalization methods to enhance background similarity and crack saliency.

With the rapid development of deep learning, concrete surface crack detection methods based on convolutional neural networks (CNNs) have gradually become mainstream. These methods significantly enhance the accuracy and robustness of crack detection by automatically learning deep features within images. The U-Net model [12], known for its successful application in medical image segmentation, is widely used for concrete surface crack detection. Qiao et al. [13] employed an improved U-Net model to identify crack widths in binary images of concrete cracks. Faster R-CNN [14] is another deep learning model commonly used for object detection; Li et al. [15] used a faster R-CNN algorithm based on VGG16 transfer learning to design a two-RTS system for drone hover precision to detect bridge surface cracks. Mask R-CNN [16] can not only detect cracks but also accurately segment crack areas. Liu et al. [17] developed an improved mask R-CNN to automatically detect and segment small cracks in asphalt pavements at the pixel level. The DeepLab model [18] captured multi-scale contextual information in images using atrous convolution. Sun et al. [19] proposed a multi-scale attention module in the decoder of DeepLabv3+, generating attention masks and dynamically allocating weights between deep and shallow feature maps. The residual network (ResNet) [20] solved the problem of gradient disappearance in deep networks by introducing residual connections. Fan et al. [21] proposed a novel deep residual convolutional neural network called Parallel ResNet and designed a pavement crack detection and recognition system. DenseNet [22] improved feature reusability through dense connections between layers. López Droguett et al. [23] proposed a DenseNet architecture that achieves better crack detection perfor-

mance than standard algorithms with only a small number of parameters. The VGG model [24] had excellent performance in feature extraction due to its deep structure. Que et al. [25] proposed an improved VGG model for crack classification. EfficientNet [26] used a composite scaling strategy to improve performance while maintaining model efficiency. Satheesh et al. [27] used EfficientNet for crack segmentation. The transformer model [28] has been gradually applied to image-processing tasks due to its success in natural language processing. Shamsabadi et al. [29] proposed a framework based on the vision transformer for crack detection on asphalt and concrete surfaces. Generative adversarial networks (GANs) [30] generate high-quality images through adversarial training. Sekar et al. [31] introduced a conditional prediction crack GAN (CFC-GAN) model to detect pavement crack images under various conditions.

In summary, the automation and intelligence of concrete crack detection can significantly reduce labor costs and save time while avoiding errors inherent in traditional manual inspections. For large-scale engineering projects and infrastructure, such as bridges, dams, and high-rise buildings, intelligent crack detection technology provides real-time and accurate safety assessments. This, in turn, helps prevent potential risks and avoids major accidents and economic losses caused by crack propagation. Among the intelligence algorithms, TransUNet has demonstrated excellent performance in image segmentation and has been applied to crack detection, but it still faces several critical challenges, such as the limitation in generalization capability under complex backgrounds, the difficulty in capturing fine-grained crack features, and the lack in effective feature fusion strategy between deep and shallow representations. These issues limit the applicability of TransUNet in complex scenarios.

So, this study aims to develop an advanced model that improves the accuracy and completeness of crack detection under complex environmental conditions. An improved TransUNet-based model, AG-TransUNet, is proposed. The model introduces an adaptive multi-head self-attention mechanism, which enables the model to dynamically adjust its feature extraction based on the specific features of the input image, thereby enhancing crack localization in diverse scenarios. Additionally, a gated mechanism-based decoding module (GRU-T) is designed to effectively preserve essential image features and reduce information loss during the upsampling process, leading to improved crack detection precision. To further quantify crack severity, an orthogonal skeleton-based crack width estimation method is proposed, enabling precise crack width measurement, which serves as a crucial reference for structural safety assessment. The model provides a more robust and accurate crack detection solution for engineering applications.

2. Related Principles

2.1. Overview of the TransUNet Algorithm

TransUNet [32] is a hybrid deep learning model combining a transformer and U-Net, which was originally proposed for medical image segmentation tasks. This algorithm aims to combine the global feature extraction capability of the transformer and the multi-scale context information fusion capability of U-Net to improve the accuracy and robustness of image segmentation.

As shown in Figure 1, TransUNet embeds the transformer module into the encoder part of the U-Net, creating a novel segmentation network architecture. The encoder in TransUNet uses convolutional operations to extract shallow features from the input image. These extracted features are then fed into the transformer encoder for further processing. The transformer encoder employs an attention mechanism to capture global features within the image, outputting feature representations enriched with contextual information. After processing through the transformer, the feature maps are passed to the U-Net decoder,

which performs step-by-step upsampling and feature fusion to generate high-resolution segmentation results.

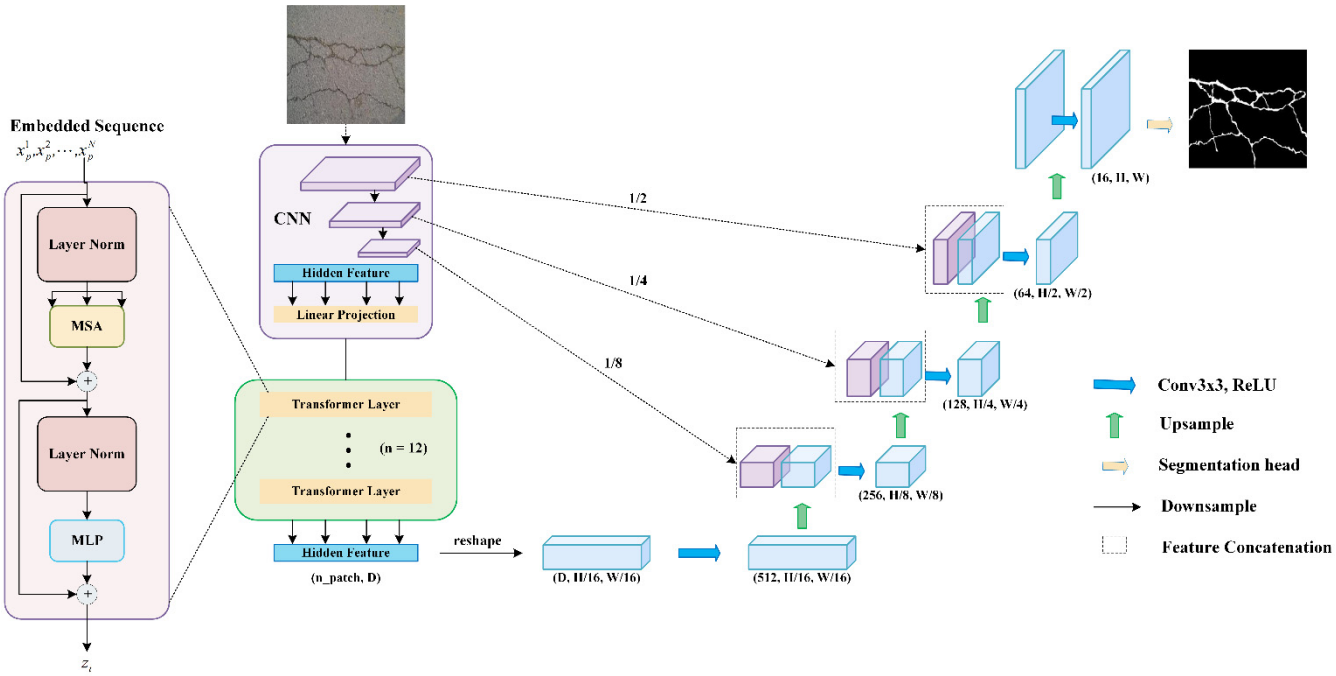

Figure 1. Structure of TransUNet.

The main advantage of TransUNet is its ability to utilize local and global feature information simultaneously, which makes it highly effective when dealing with targets of complex structures and irregular shapes.

2.2. Limitations of the TransUNet Algorithm

The existing TransUNet model has certain limitations when processing images with high noise or complex backgrounds. These limitations mainly manifest in two areas. Firstly, although the attention mechanism in TransUNet can capture the global features of an image, its fixed attention head configuration lacks sufficient flexibility in processing images with different feature patterns. This rigid allocation of attention heads may not adequately adapt to the feature changes in various complex scenes. As a result, the model cannot accurately focus on small targets such as cracks or ignore the interference information in the background, making the segmentation effect unsatisfactory in some scenarios.

Secondly, in the decoding phase, TransUNet relies on simple convolution operations for feature reconstruction. Although this approach can partially restore spatial resolution, it does not fully exploit the rich feature information passed from the encoder to the decoder. This limitation is particularly problematic when dealing with images that involve complex structures and intricate details, potentially leading to information loss or blurring. This information loss hinders the effective fusion of high-level semantic information with low-level detail, potentially causing the model to produce inaccuracies in boundary delineation and detail processing, ultimately affecting the segmentation accuracy and robustness. In summary, the current TransUNet model's inflexible attention mechanism and the simplistic feature reconstruction process limit its performance in complex environments, particularly when precise detection of fine structures like cracks is crucial.

3. Improvements to the TransUNet Algorithm

This paper, therefore, proposes an improved TransUNet crack segmentation algorithm to enhance the model's adaptability and segmentation accuracy in complex scenarios.

Figure 2 shows the structure of the improved TransUNet algorithm. The enhancement focuses on the attention mechanism within TransUNet, replacing the traditional fixed attention configuration with an adaptive multi-head self-attention mechanism. This adaptive mechanism dynamically adjusts the number and distribution of attention heads based on the characteristics of the input image. This adjustment enables the model to more accurately capture key information related to cracks in diverse image scenes, reducing its sensitivity to background noise and improving the reliability of the segmentation results. In addition, to address the issue of information loss during the upsampling process, a gated mechanism-based decoding module (GRU-T) is introduced into TransUNet's decoder. As a recurrent neural network unit, GRU-T utilizes its built-in gating mechanism to effectively control the flow of information, retaining critical feature information while minimizing redundancy and noise.

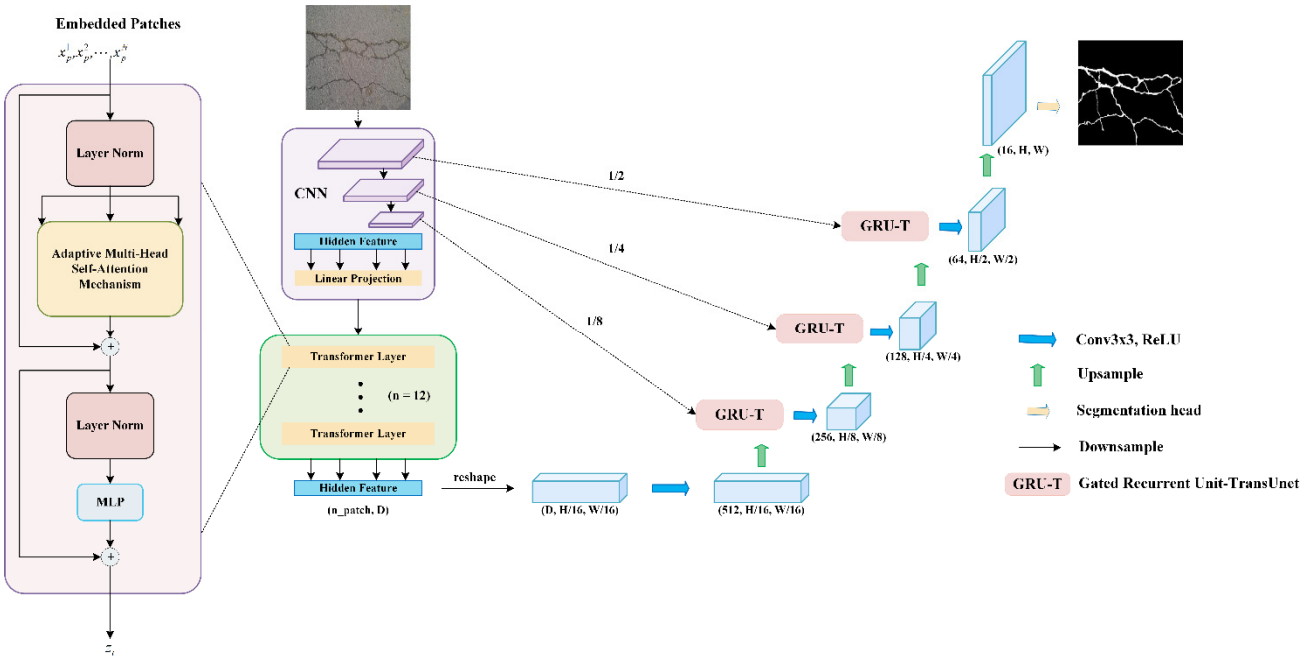

Figure 2. Structure of the AG-TransUNet.

3.1. Adaptive Multi-Head Self-Attention Mechanism

The adaptive multi-head self-attention mechanism is an enhanced attention mechanism designed to improve adaptability across various image scenarios. Unlike the traditional multi-head self-attention mechanism, this adaptive approach dynamically adjusts the number and weight distribution of attention heads, allowing the model to flexibly optimize attention configurations based on the features of the input image. This capability enables the model to capture key information more effectively, especially in complex backgrounds or high-noise environments, thereby improving segmentation accuracy and reducing sensitivity to irrelevant information. The adaptive multi-head self-attention mechanism extracts high-dimensional features through convolutional layers and then dynamically adjusts the number of attention heads based on the complexity of the input features and assigns the most suitable weights to each attention head. The final output is a fused attention map containing the crucial information. Figure 3 is the structure of the adaptive multi-head self-attention mechanism.

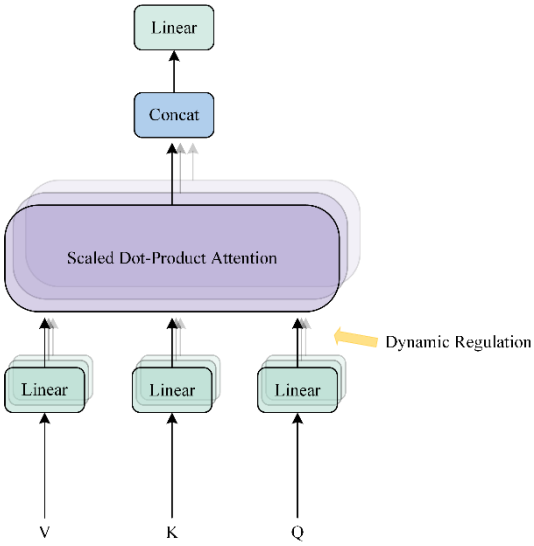

Figure 3. Structure of the adaptive multi-head self-attention mechanism.

3.2. GRU-T

The gated recurrent unit (GRU) [33] is a variant of the recurrent neural network (RNN) designed to solve the problem of gradient vanishing and exploding gradients in standard RNNs. By introducing a gating mechanism, GRU enables the network to better capture dependencies in long sequences while reducing computational complexity.

In TransUNet, the decoding process relies on simple convolution operations for feature reconstruction. While this approach can partially restore spatial resolution, it fails to fully exploit and utilize the rich feature information transmitted from the encoder to the decoder. Especially when dealing with images involving complex structures and abundant details, it may lead to loss or blurring of information. The GRU-T module is therefore introduced to enhance the recovery of crack details during the upsampling process and improve the completeness of crack segmentation. The structure of GRU-T is shown in Figure 4. It leverages GRU to strengthen shallow and deep features and is then fused with the downsampled features. This approach effectively preserves critical crack information.

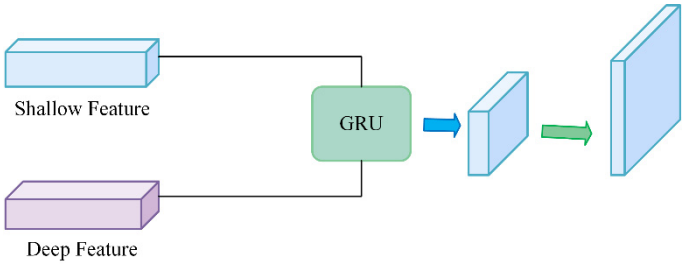

Figure 4. Structure of the GRU-T module.

The structure of the GRU is shown in Figure 5. GRU uses two gating mechanisms (reset gate and update gate) to flexibly control the flow of information. It can make effective use of past information and quickly update the current hidden state as needed.

The principles of the GRU are as follows:

The update gate determines the extent to which historical information and current information are used to update the current hidden state. At time t, the update gate is as follows:

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t]) \quad (1)$$

where z_t is the gated update signal, the size of z_t determines the degree of memory of the candidate hidden state, h_{t-1} is the historical hidden state, x_t represents the input data at time t, W_z is the weight matrix, and σ is the sigmoid function.

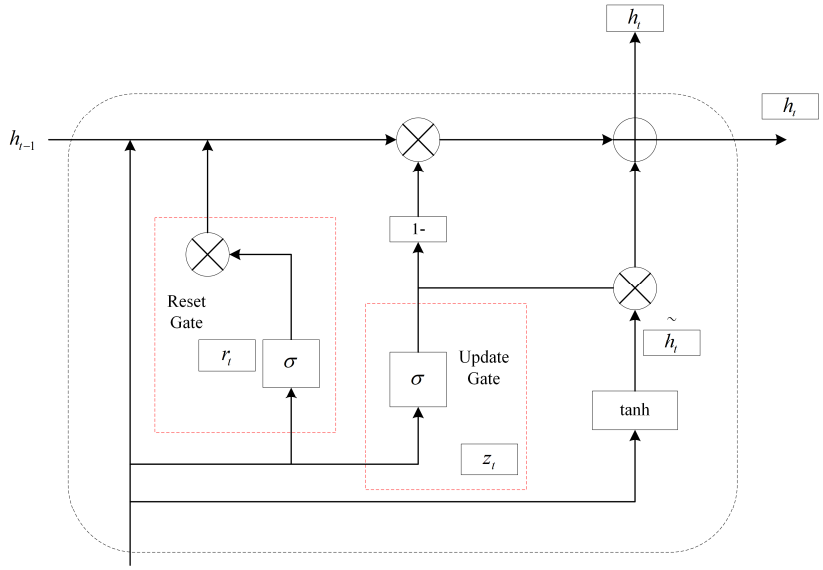

Figure 5. Structure of the GRU.

The reset gate determines how much historical information is retained. The reset gate at time t is as follows:

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t]) \tag{2}$$

where r_t is the reset signal; the larger the reset signal value, the more historical information needs to be remembered, and W_r is the weight matrix.

Under the action of the update gate z_t and reset gate r_t, the candidate hidden state and hidden output state h_t at the current time can be updated as follows:

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t \tag{3}$$

Among them, the candidate hidden states are

$$\tilde{h}_t = tanh(W \cdot [r_t * h_{t-1}, x_t]) \tag{4}$$

In this equation, *tanh* is the hyperbolic tangent function, and the candidate hidden state is responsible for fusing the information features of the input data and the historical data. This operation is related to the reset signal r_t obtained by the reset gate. h_t represents the final cell state at the current time, which includes two processes: forgetting and remembering. The product of $(1 - z_t)$ and the hidden state h_{t-1} at the previous time represents the forgetting process. The closer z_t is to 1, the more information will be forgotten at the previous time. The product of z_t and the candidate hidden state represents the remembering process, and the size of z_t determines the memory degree of the candidate hidden state, that is, how much of the previous hidden state is retained.

3.3. Orthogonal Skeleton Method

The orthogonal skeleton method is a technique for analyzing geometric features by extracting the object's skeleton and measuring its distance to the edges. It has strong robustness to image noise and edge irregularities because the skeleton is typically located

in the central region of the object, and it is less affected by edge noise. This method is especially suitable for handling objects with complex shapes.

The method first preprocesses the image (denoising, grayscale conversion, and binarization). Then, a thinning algorithm is employed to extract the skeleton of the object, which represents the geometric center of the object. The orthogonal direction (perpendicular to the skeleton) for each skeleton point is calculated, and the distance from the skeleton point to the object's edge is measured along this direction. These orthogonal distances are used to analyze the geometric characteristics of the object, such as its width and shape, especially for the accurate calculation of crack widths.

4. Experiments and Results

4.1. Dataset

In this study, two datasets are used for model training: the CFD dataset and the concrete crack dataset. The CFD public dataset consists of 118 crack images with a resolution of 480 × 320 pixels, containing noise such as water stains and shadows. The concrete crack dataset is a self-made dataset obtained by capturing images of concrete surface cracks using a camera. It contains 332 images with a resolution of 224 × 224 pixels, designed to test the model's performance in practical engineering applications.

Given that both datasets are relatively small and the proposed model requires input images of 256 × 256 pixels, data augmentation techniques are applied. These techniques include random cropping, brightness adjustment, contrast adjustment, and angle rotation. As a result, the CFD dataset was expanded to 3658 images, and the concrete crack dataset was expanded to 4700 images. Each dataset was randomly divided into a training set and a test set in an 8:1 ratio. When training data are limited, the model may overly rely on the small set of available samples, leading to poor performance on new, unseen data. Data augmentation effectively increases the size and diversity of the training dataset, mitigating the risk of overfitting during the training process.

4.2. Experimental Setup and Evaluation Metrics

The proposed model is developed in Python 3.10, with the open-source deep learning framework PyTorch serving as the network framework. The training process is accelerated by CUDA 11.8. The hardware environment for model testing includes an Intel® Xeon® Platinum 8375C CPU @ 2.90 GHz and an NVIDIA RTX 4090 GPU with 24 GB of VRAM.

During training, the stochastic gradient descent (SGD) optimizer is used. The model is trained for 100 epochs with a batch size of 16 and an initial learning rate of 0.01.

The evaluation metrics used in this study include the F1-score, precision (P), recall (R), and intersection over union (IoU). The P and R are the basic metrics, while the F1-score and IoU, which are derived from the P and R, are used as the final evaluation indicators.

Precision is the ratio of correctly predicted positive samples to the total number of samples predicted as positive:

$$Precision = \frac{TP}{TP + FP}, \tag{5}$$

Recall is calculated as the proportion of all actual targets that are correctly predicted:

$$Recall = \frac{TP}{TP + FN}, \tag{6}$$

Here, TP is the number of correctly detected targets, FP is the number of incorrectly detected targets, and FN is the number of missed targets among the actual correct targets.

The F1-score takes precision and recall into account, providing a more comprehensive reflection of the overall performance of the network. It is calculated as the harmonic mean of these two metrics, as shown in (7):

$$F1 = 2\frac{Precision \times Recall}{Precision + Recall}, \quad (7)$$

4.3. Experimental Results and Analysis

4.3.1. Comparison of Ablation Experiments

To verify the effectiveness of incorporating the adaptive multi-head self-attention mechanism and the GRU-T module into the transformer's encoder in the proposed algorithm, an ablation experiment was conducted on the CFD dataset, comparing the modified model with the original TransUNet under the same conditions. This study was divided into four groups: Experiment 0 used the original TransUNet, Experiment 1 introduced the adaptive multi-head self-attention mechanism into the original TransUNet, Experiment 2 improved the decoder of the original TransUNet by using the GRU-T module, and Experiment 3 combined the improvements from Experiments 1 and 2 into the original TransUNet.

Table 1 presents the comparison of the ablation experiment results. Compared to Experiment 0, Experiment 1 achieves an increase of 2.51% in F1-score and 2.84% in IoU, demonstrating that the introduction of the adaptive multi-head self-attention mechanism significantly enhances segmentation performance. However, the processing time increases from 5899 s to 6513 s, indicating that while performance improves, the self-attention mechanism also adds computational complexity, especially with its higher computational cost. When comparing Experiment 0 with Experiment 2, F1-score increases by 4.82%, and IoU improves by 2.77%, validating the effectiveness of the GRU-T module. The module integrates the temporal processing capabilities of gated recurrent units (GRU), enabling better fusion of deep feature representations and shallow edge details. By reducing information loss during the decoding process, the GRU-T module enhances segmentation accuracy, particularly in detecting elongated and irregular cracks. Despite these improvements, the processing time increases from 5899 s to 6309 s, indicating the additional computational burden is also introduced by the GRU module. Compared to Experiment 0, Experiment 3 shows an 8.28% improvement in F1-score and a 3.54% increase in IoU, showing that the combination of both enhancements significantly improves crack segmentation performance. Despite the introduction of both modules, the processing time decreases from 6513 s to 5667 s. The results indicate that the combination of the adaptive multi-head self-attention mechanism and the GRU-T module leads to optimization of the computation efficiency, reduction of the redundant computations, and improvement of the processing speed. Furthermore, the integration of the two mechanisms fully leverages the feature extraction capabilities of the transformer encoder while optimizing the information reconstruction ability of the decoder, making the crack detection model more stable and reliable.

Table 1. Comparison of ablation experiment results.

No.	Attention Mechanism	GRU-T	F1-Score/%	IoU/%	Time/s
0			80.36	73.52	5899
1	√		82.87	76.36	6513
2		√	85.18	76.29	6309
3	√	√	88.64	77.06	5667

By improving the attention mechanism in TransUNet to an adaptive multi-head self-attention mechanism, the key information of cracks can be captured more effectively, and

the robustness of the model across diverse scenarios is also enhanced. The eight-head attention mechanism effectively balances global context modeling and local feature extraction, capturing multi-scale dependencies while reducing false detections in complex backgrounds. Although the introduction of the adaptive multi-head self-attention mechanism increases computational overhead due to the complexity of feature extraction and attention weight calculations, AG-TransUNet incorporates multiple optimization strategies to ensure that the additional computational cost remains within a manageable range.

Compared to the standard attention mechanism in TransUNet, the adaptive self-attention mechanism dynamically adjusts attention weight distributions, improving the model's ability to focus on crack-related features while reducing redundant information. To prevent excessive computational costs, the number of attention heads is set to eight, balancing global and local feature dependencies while avoiding exponential parameter growth. Furthermore, efficient attention computation methods are employed to reduce redundant calculations and optimize GPU memory utilization, ensuring a rational allocation of computational resources. At the same time, the GRU-T decoding module, designed for the upsampling stage, enables better integration of deep and shallow features, reducing information loss during feature reconstruction and enhancing segmentation accuracy. A two-layer GRU with 128 hidden units is employed, providing an optimal balance between computational efficiency and feature refinement. The GRU-based sequential modeling smooths crack boundaries and mitigates segmentation tasks, while the two-layer configuration outperforms single-layer designs without introducing unnecessary complexity. Ablation experiments show that AG-TransUNet achieves optimal performance in terms of F1-score and IoU. So, these architectural improvements significantly improve the model's ability to segment cracks on concrete surfaces with better accuracy and robustness.

4.3.2. Evaluation of Practical Application

1. CFD Dataset Experiment

To validate the performance of the improved algorithm in the segmentation model, experiments were conducted on the CFD dataset. The quantitative results of this dataset under different models are presented in Table 2. The table shows that AG-TransUNet achieved a precision of 91.26%, an F1-score of 88.64%, and an IoU of 77.06%. Compared with the original TransUNet model, the improvements were 4.05%, 2.59%, and 0.36%, respectively. AG-TransUNet demonstrated superior performance in terms of precision, F1-score, and IoU, resulting in better segmentation outcomes.

Table 2. Comparison of evaluation metrics for different models on the CFD dataset.

Model	Precision/%	Recall/%	F1-Score/%	IoU/%
SegFormer	70.32	74.05	72.19	77.63
U-Net	76.28	72.78	73.64	72.59
FCN	77.23	71.03	72.56	71.36
PSPNet	77.29	72.51	74.69	72.04
TransUNet	87.21	85.69	86.05	76.70
AG-TransUNet	91.26	87.87	88.64	77.06

To further illustrate the segmentation performance of the improved model, Figure 6 shows the segmentation results of three images from the test set using different models. The comparison indicates that algorithms like U-Net perform poorly in crack segmentation, often resulting in missed detections. By contrast, AG-TransUNet provides a better segmentation effect, with fewer cases of missed or false detections.

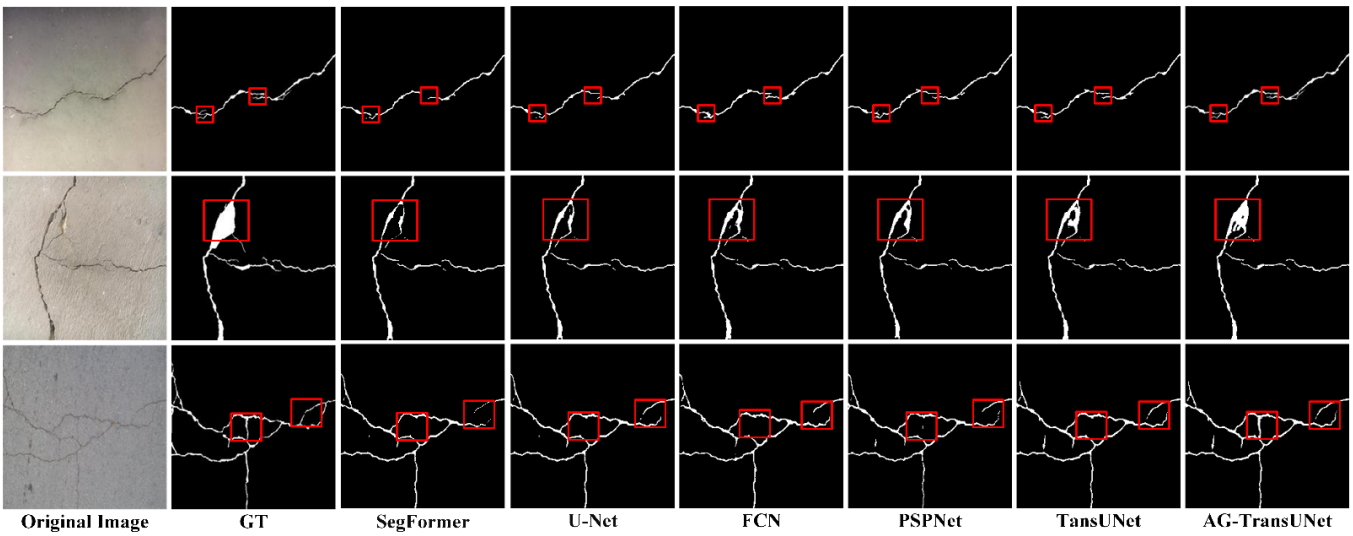

Figure 6. Comparison of visualization results on the CFD dataset.

2. Concrete Crack Dataset Experiment

To further and more comprehensively validate the effectiveness of the improved model, experiments were also conducted on the concrete crack dataset. Table 3 presents the quantitative analysis results of different models on this dataset. It can be found that AG-TransUNet exhibits outstanding performance across all metrics, achieving a precision of 86.48%, an F1-score of 87.11%, and an IoU of 78.05%. Compared with the original TransUNet model, AG-TransUNet increased by 2.21%, 5.63%, and 9.07% in these three metrics, respectively. These results showed that AG-TransUNet is significantly superior to other segmentation models in terms of precision, F1-score, and IoU and achieved more accurate crack segmentation.

Table 3. Comparison of evaluation metrics for different models on the concrete crack dataset.

Model	Precision/%	Recall/%	F1-Score/%	IoU/%
SegFormer	67.71	76.59	71.85	77.70
U-Net	65.89	65.44	64.29	60.41
FCN	58.80	68.32	63.57	58.34
PSPNet	53.47	63.28	55.19	40.39
TransUNet	84.27	78.30	81.48	68.98
AG-TransUNet	86.48	87.63	87.11	78.05

To visually demonstrate the segmentation performance of the improved model, three representative images from the test set were selected and compared with different models, as shown in Figure 7. The comparison revealed that some traditional models exhibit noticeable shortcomings in crack segmentation, often resulting in missed detections. By contrast, AG-TransUNet was significantly superior to other models in segmentation accuracy, and cases of missing and false detection were few, fully proving its excellent performance in crack detection.

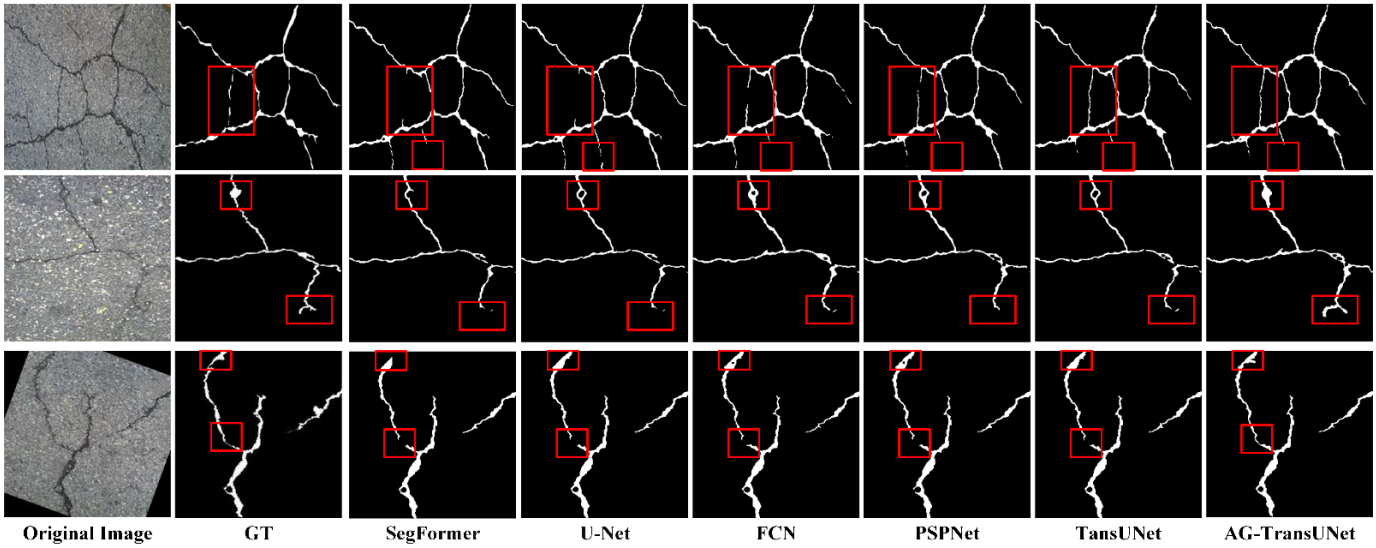

Figure 7. Comparison of visualization results on the concrete crack dataset.

4.3.3. Orthogonal Skeleton Method and Crack Width

In crack evaluation and detection, the precise measurement of crack width is a crucial step in assessing structural safety and durability. Crack width not only affects the load-bearing capacity of the structure but also reflects the extent of structural damage. Accurately detecting crack width is, therefore, essential for the early identification of potential issues and the implementation of appropriate maintenance measures. To address this problem, after accurately segmenting the cracks using the AG-TransUNet algorithm, this study employs the orthogonal skeleton method to calculate the crack width accurately. Figure 8 shows the central line of the crack extracted by the orthogonal skeleton method.

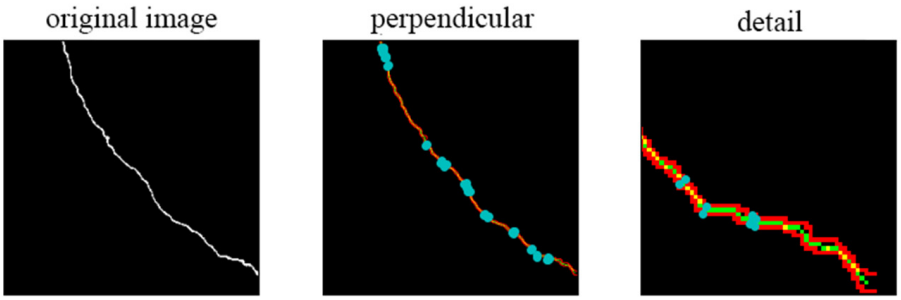

Figure 8. Extraction of the crack centerline.

To further validate the accuracy of the proposed model in crack width calculation, this study selected 10 concrete surface cracks from the concrete crack dataset as experimental subjects. These cracks cover a variety of typical shapes and sizes and have a certain representativeness. At the same time, to ensure the accuracy of the measurement and the consistency of the results, a uniform shooting height is used for all crack images to ensure that the distance between the camera and the crack surface is consistent. This height helps to avoid image distortion or scale error caused by different shooting angles or distances, thus improving the accuracy of crack width calculation. Table 4 shows the calculation results and errors of different crack widths.

As shown in Table 4, the proposed model demonstrates high accuracy and consistency in calculating crack width. The experimental results show that the model can accurately measure crack width under the influence of complex fracture morphology, verifying its application potential in practical engineering.

Table 4. Analysis of crack width calculation results.

Crack No.	Calculated Width/mm	Actual Width/mm	Error/mm	Relative Error/%
1	3.44	3.26	0.18	5.52
2	6.26	5.97	0.29	4.86
3	4.98	4.72	0.26	5.51
4	4.95	5.00	0.05	1.00
5	4.65	4.87	0.21	4.31
6	13.08	12.95	0.13	1.00
7	12.71	13.06	0.35	2.67
8	16.35	17.21	0.86	4.99
9	22.89	21.75	1.14	5.24
10	23.52	22.69	0.83	3.65

However, compared with the actual crack width, the error sources of the model are primarily attributed to image quality, crack edge complexity, and the inherent limitations of the model itself. Low-resolution images, uneven lighting, and noise interference may degrade the segmentation accuracy, leading to errors. Additionally, irregular or blurred crack edges, the presence of intersecting cracks, and abrupt width variations could also introduce errors in skeleton extraction and width estimation. Limitations of the model itself, such as the approximations in skeleton extraction and potential distortions in feature reconstruction, may also lead to minor errors. As shown in Table 4, the relative error varies with different cracks, ranging from 1.00% to 5.52%. Cracks with well-defined edges and stable widths (e.g., Crack No. 4 and Crack No. 6) have minimal error (1.00%), while cracks with irregular borders or width fluctuations (e.g., Crack No. 1 and Crack No. 9) show higher relative error (5.52% and 5.24%, respectively).

5. Conclusions

Concrete surface crack detection is a critical research topic in structural safety, particularly in the safety assessment of roads, bridges, and buildings. Traditional crack detection methods typically rely on manual inspection, which is not only time-consuming and labor-intensive but also inefficient. Consequently, the use of deep learning algorithms to automate crack detection has become a hot point of the current study. TransUNet, a segmentation algorithm that combines the transformer and U-Net, has shown significant advantages in crack detection. However, when dealing with the complexity and diversity of images in a specific environment, the generalization capability of TransUNet can be limited. For instance, in the presence of other textures, stains, or shadows on concrete surfaces, TransUNet may be prone to interference, resulting in a noticeable decline in crack detection accuracy. To address these problems, this paper proposes an improved TransUNet-based crack segmentation algorithm to overcome the shortcomings of traditional TransUNet in concrete crack detection. The main conclusions are as follows:

(1) By introducing the adaptive multi-head self-attention mechanism, this study significantly enhances the model's flexibility and accuracy in crack detection. This mechanism dynamically adjusts the number and distribution of attention heads based on the features of the input image, allowing the model to autonomously optimize the allocation of attention resources when processing images of varying complexity. This enables the precise capture of key crack information. The mechanism is particularly effective in high-noise environments and complex backgrounds, substantially reducing the probability of false positives and missed detections, thereby providing support for the accuracy and robustness of crack detection.

(2) To further improve crack segmentation performance, this study designs and implements a novel decoding module, GRU-T. This module combines the temporal sequence processing capability of the GRU with the image processing functions of a traditional decoder, enabling a more effective fusion of deep feature information with shallow detail information. The GRU-T module is particularly suited for handling crack images in complex backgrounds because it can capture fine crack features and preserve edge details, thereby enhancing segmentation accuracy. Additionally, the module shows good performance in processing elongated and narrow cracks, reducing edge discontinuities, and mitigating the impact of noise on the segmentation results.

(3) This paper proposes a crack width calculation method based on the orthogonal skeleton line method to address the limitations of traditional methods in measurement accuracy. By extracting the skeleton line of the crack and calculating the width along the orthogonal direction, this method can accurately measure the actual crack width, making it particularly suitable for cracks with complex shapes and blurred edges. Experimental results demonstrate that the application of the orthogonal skeleton line method on the dataset used in this study achieves good measurement accuracy. The method provides a reliable and efficient solution for crack width measurement in structural health monitoring.

(4) The improved model proposed in this paper demonstrates superior performance in crack detection and crack width calculation. Experiments on different datasets fully validate that the model has efficient detection ability. On the CFD dataset, AG-TransUNet outperforms the original TransUNet with a 4.05% increase in precision, a 2.59% improvement in F1-score, and a 0.36% enhancement in IoU. On the concrete crack dataset, AG-TransUNet achieves a 2.21% increase in precision, a 5.63% improvement in F1-score, and a 9.07% enhancement in IoU. Additionally, the crack width calculation method based on the orthogonal skeleton approach achieves an average error of 3.88%.

(5) Although AG-TransUNet shows better segmentation accuracy and robustness, it still has some limitations. The model's performance is affected by image quality, variations in crack morphology, and environmental conditions, which may affect its generalization in different scenarios. Future research will focus on further optimizing the design of the adaptive multi-head self-attention mechanism and the GRU-T module, particularly to enhance the model's segmentation accuracy in complex environments and reduce the error of crack width calculation.

Author Contributions: Conceptualization, X.D. and J.D.; methodology, Y.L.; software, Y.L.; validation, X.D and J.D.; formal analysis, Y.L.; investigation, Y.L. and X.D.; resources, X.D. and J.D.; data curation, Y.L. and X.D.; writing—original draft preparation, Y.L.; writing—review and editing, Y.L.; visualization, Y.L.; supervision, X.D.; project administration, X.D. and J.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by the National Natural Science Foundation of China (Grant No. 52368032, 51808272), the China Postdoctoral Science Foundation (Grant No. 2023M741455), the Tianyou Youth Talent Lift Program of Lanzhou Jiaotong University, the Gansu Province Youth Talent Support Project (Grant No. GXH20210611-10), and in part by the Natural Science Foundation of Gansu Province (Grant No. 23JRRA889) and the Innovation Fund Project of Colleges and Universities in Gansu Province (Grant No. 2024B-057).

Data Availability Statement: Part of the dataset used in this article is a public dataset, which can be found on the Internet, and the dataset we created can be requested from the corresponding author upon reasonable request.

Acknowledgments: The authors would like to thank the technical team of the Key Laboratory of. Opto-Electronic Technology and Intelligent Control of the Ministry of Education, Lanzhou Jiaotong University, and the National and Provincial Joint Engineering Laboratory of Road & Bridge Disaster Prevention and Control, Lanzhou Jiaotong University, for their technical support.

Conflicts of Interest: Author Jinpeng Dai was employed by the company Jiangsu Research Institute of Building Science Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Al-Zu'bi, M.; Fan, M.; Al Rjoub, Y.; Ashteyat, A.; Al-Kheetan, M.J.; Anguilano, L. The effect of length and inclination of carbon fiber reinforced polymer laminates on shear capacity of near-surface mounted retrofitted reinforced concrete beams. *Struct. Concr.* **2021**, *22*, 3677–3691. [CrossRef]
2. Qu, Z.; Chen, W.; Wang, S.Y.; Yi, T.M.; Liu, L. A Crack Detection Algorithm for Concrete Pavement Based on Attention Mechanism and Multi-Features Fusion. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 11710–11719. [CrossRef]
3. Canny, J. A Computational Approach to Edge Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **1986**, *PAMI-8*, 679–698. [CrossRef]
4. Zhang, W.Y.; Zhang, Z.J.; Qi, D.P.; Liu, Y. Automatic Crack Detection and Classification Method for Subway Tunnel Safety Monitoring. *Sensors* **2014**, *14*, 19307–19328. [CrossRef]
5. Otsu, N. A Threshold Selection Method from Gray-Level Histograms. *IEEE Trans. Syst. Man Cybern.* **1979**, *9*, 62–66. [CrossRef]
6. Vivekananthan, V.; Vignesh, R.; Vasanthaseelan, S.; Joel, E.; Kumar, K.S. Concrete bridge crack detection by image processing technique by using the improved OTSU method. *Mater. Today* **2023**, *74*, 1002–1007. [CrossRef]
7. Oron, S.; Dekel, T.; Xue, T.; William, T.F.; Avidan, S. Best-Buddies Similarity-Robust Template Matching Using Mutual Nearest Neighbors. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 1799–1813. [CrossRef] [PubMed]
8. Chen, W.; Zhang, J. Efficient and lightweight monitoring network for cracks in complex background regions based on adaptive perception. *Automat. Constr.* **2024**, *166*, 105614. [CrossRef]
9. Gharehbaghi, V.; Farsangi, N.E.; Yang, T.Y.; Noori, M.; Kontoni, D.-P.N. A Novel Computer-Vision Approach Assisted by 2D-Wavelet Transform and Locality Sensitive Discriminant Analysis for Concrete Crack Detection. *Sensors* **2022**, *22*, 8986. [CrossRef]
10. Arya, D.; Ghosh, S.K.; Toshniwal, D. Automatic Recognition of Road Cracks Using Gray-Level Co-occurrence Matrix and Machine Learning. In Proceedings of the 2022 International Conference on Machine Intelligence and Signal Processing, Allahabad, India, 12–14 March 2022.
11. Liu, J.; Zhao, Z.; Lv, C.S.; Ding, Y.F.; Chang, H.L.; Xie, Q.Y. An image enhancement algorithm to improve road tunnel crack transfer detection. *Constr. Build. Mater.* **2022**, *348*, 128583. [CrossRef]
12. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the 2015 Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015.
13. Qiao, W.T.; Zhang, H.W.; Zhu, F.; Wu, Q.D. A crack identification method for concrete structures using improved U-Net convolutional neural networks. *Math. Probl. Eng.* **2021**, *2021*, 6654996. [CrossRef]
14. Ren, S.Q.; He, K.M.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
15. Li, R.X.; Yu, J.Y.; Li, F.; Yang, R.T.; Wang, Y.D. Automatic bridge crack detection using Unmanned aerial vehicle and Faster R-CNN. *Constr. Build. Mater.* **2023**, *362*, 129659. [CrossRef]
16. He, K.M.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the 2017 IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
17. Liu, Z.; Yeoh, J.K.W.; Gu, X.Y.; Dong, Q.; Chen, Y.H.; Wu, W.X.; Wang, L.T.; Wang, D.Y. Automatic pixel-level detection of vertical cracks in asphalt pavement based on GPR investigation and improved mask R-CNN. *Automat. Constr.* **2023**, *146*, 104689. [CrossRef]
18. Chen, L.C.; Papandreou, G.; Kokkinos, L.; Murphy, K.; Yuille, A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 834–848. [CrossRef] [PubMed]
19. Sun, X.Z.; Xie, Y.C.; Jiang, L.M.; Cao, Y.; Liu, B.Y. DMA-Net: DeepLab with Multi-Scale Attention for Pavement Crack Segmentation. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 18392–18403. [CrossRef]
20. He, K.M.; Zhang, X.Y.; Ren, S.Q.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.

21. Fan, Z.; Lin, H.B.; Li, C.; Sun, J.; Bruno, S.; Loprencipe, G. Use of Parallel ResNet for High-Performance Pavement Crack Detection and Measurement. *Sustainability* **2022**, *14*, 1825. [CrossRef]
22. Huang, G.; Liu, Z.; Laurens, V.D.M.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
23. López Droguett, E.; Tapia, J.; Yáñez, C.; Boroschek, R. Semantic segmentation model for crack images from concrete bridges for mobile devices. *Proc. Inst. Mech. Eng. Part O J. Risk Reliab.* **2020**, *236*, 570–583. [CrossRef]
24. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556. [CrossRef]
25. Que, Y.; Dai, Y.; Ji, X.; Leung, A.K.; Chen, Z.; Jiang, Z.L.; Tang, Y.C. Automatic classification of asphalt pavement cracks using a novel integrated generative adversarial network and improved VGG model. *Eng. Struct.* **2023**, *277*, 115406. [CrossRef]
26. Tan, M.X.; Le, Q.V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *arXiv* **2019**, arXiv:1905.11946. [CrossRef]
27. Satheesh, K.G.; Narender, C.; Selvan, S.T.; Rajakum, V. Automatic Detection of Road Cracks using EfficientNet with Residual U-Net-based Segmentation and YOLOv5-based Detection. *Int. J. Recent Innov. Trends Comput. Commun.* **2023**, *11*, 17762.
28. Vaswani, A.; Shazeer, N.; Paramar, N.; Uszkoreit, J.; Jones, L.; N.Gomea, A.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762. [CrossRef]
29. Shamsabadi, E.A.; Xu, C.; Rao, A.S.; Nguyen, T.; Ngo, T.; Dias-da-Costa, D. Vision transformer-based autonomous crack detection on asphalt and concrete surfaces. *Automat. Constr.* **2022**, *140*, 104316. [CrossRef]
30. Moez, K. Generative Adversarial Networks. In Proceedings of the 2023 International Conference on Computing Communication and Networking Technologies, Delhi, India, 6–8 July 2023.
31. Seker, A.; Perumal, V. CFC-GAN: Forecasting Road Surface Crack Using Forecasted Crack Generative Adversarial Network. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 21378–21391. [CrossRef]
32. Chen, J.N.; Lu, Y.Y.; Yu, Q.H.; Luo, X.D.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y.Y. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. *arXiv* **2021**, arXiv:2102.04306. [CrossRef]
33. Chung, J.; Gulcehre, C.; Cho, K.H.; Bengio, Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *arXiv* **2014**, arXiv:1412.3555. [CrossRef]