electronics

MDPI

Article

Feature Sparse Choosing VIT Model for Efficient Concrete Crack Segmentation in Portable Crack Measuring Devices

Xiaohu Zhang, Haifeng Huang * and Meng Cai

School of Electronic and Communication Engineering, Sun Yat-sen University, Guangzhou 510006, China; zhangxh79@mail2.sysu.edu.cn (X.Z.)
* Correspondence: huanghaifeng@mail.sysu.edu.cn

Abstract: Concrete crack measurement is important for concrete buildings. Deep learning-based segmentation methods have achieved state-of-art results. However, the model size of these models is extremely large which is impossible to use in portable crack measuring devices. To address this problem, a light-weight concrete crack segmentation model based on the Feature Sparse Choosing VIT (LTNet) is proposed by us. In our proposed model, a Feature Sparse Choosing VIT (FSVIT) is used to reduce computational complexity in VIT as well as reducing the number of channels for crack features. In addition, a Feature Channel Selecting Module (FCSM) is proposed by us to reduce channel features as well as suppressing the influence of interfering features. Finally, Depthwise Separable Convolutions are used to substitute traditional convolutions for further reducing computational complexity. As a result, the model size of our LTNet is extremely small. Experimental results show that our LTNet could achieve an accuracy of 0.887, 0.817 and 0.693, and achieve a recall of 0.882, 0.805 and 0.681 on three datasets, respectively, which is 3–8% higher than current mainstream algorithms. However, the model size of our LTNet is only 2 M.

Keywords: artificial intelligence; machine learning; deep neural network; image segmentation application; crack segmentation

Citation: Zhang, X.; Huang, H.; Cai, M. Feature Sparse Choosing VIT Model for Efficient Concrete Crack Segmentation in Portable Crack Measuring Devices. *Electronics* **2024**, *13*, 1641. https://doi.org/10.3390/electronics13091641

Academic Editor: Daniel Gutiérrez Reina

Received: 5 April 2024
Revised: 20 April 2024
Accepted: 21 April 2024
Published: 25 April 2024

1. Introduction

Concrete cracks are a common problem in building structures. There are many factors in the real environment that can cause cracks. For example, the load of heavy vehicles and high-speed driving accelerates the aging of the road surface. Changing humidity and temperature causes the softening, expansion and contraction of the concrete material of buildings. If not discovered and repaired in time, these cracks might further expand, causing building damage or safety hazards. Therefore, it is necessary to measure cracks in time so that they can be repaired immediately. However, traditional manual measurements of cracks are inefficient and costly, which is not suitable for large-scale measurements. Manual measurement of cracks usually uses a thin line or silk thread to divide the crack into several small segments, and uses a ruler or measuring instrument to measure the length of each segment of the thin line. These lengths are then added up to obtain the total length of the crack. With the development of computer vision technology, scholars have proposed using image segmentation methods to measure cracks.

Traditional computer image processing technologies for crack measurements usually include the following parts: (1) using self-designed digital filters such as Gaussian filter, Canny filter [1] and Gabor filter [2] to perform edge detection and extract cracks; (2) using wavelet transform for image de-noising and crack feature extraction [3,4]; (3) using minimum path selection (MPS) to capture complex curves containing closed loops and multiple branches to complete crack measurements [5]. However, due to the existence of sharp edges and complex crack feature structures, traditional computer image processing technology is easy to incorrectly split cracks in images. In addition, due to differences in

materials and construction techniques, the texture characteristics of different concrete are different, which makes these traditional technologies that rely on manual experience lack generalization and robustness and perform poorly in different scenarios.

With the rapid development of machine learning, some machine learning algorithms are also widely used in crack measurement fields [6–9]. For example, Oliveira [10] proposed a crack measurement algorithm based on entropy and image dynamic threshold, applying dynamic threshold to identify dark pixels of the image and segmenting the image into non-overlapping blocks for entropy calculation. Pan et al. [11] used the support vector machine (SVM) algorithm to identify pavement cracks on multispectral images. In addition, Shi et al. [12] applied integral channel features to define the markers that constitute cracks, introducing random structural forests and generating a high-performance crack detector capable of measuring cracks. Sheng et al. [13] used a gradient-boosted decision tree to build the crack segmentation model to measure cracks. However, although these methods could achieve good results, they require a large amount of prior knowledge and are highly dependent on specific crack features.

With the introduction of convolutional neural networks (CNN), deep learning-based models have made amazing achievements in object detection, image segmentation and image generation fields. These deep neural networks could automatically extract high-level abstract features from images; thus, researchers are gradually using deep learning models to deal with crack detection or segmentation tasks. For example, Qu et al. [14] modified the output of the FC2 layer of the LeNet-5 model, and horizontally scaled the network model to classify images, and then used the optimized VGG16 model for crack detection. Chaiyasarn et al. [15] combined a convolutional neural network (CNN) and a fully convolutional network (FCN) to segment cracks at the pixel level. Wang et al. [16] combined a fully convolutional network (FCN)) and wavelet transform structured forest (SFW) to segment cracks. In addition, researchers have noticed the excellent performance of UNet in the field of image segmentation. Therefore, scholars have improved models based on UNet to segment cracks for crack measurements. He et al. [17] proposed a CrackHAM model based on U-Net architecture, he designed a HASPP module and introduced a dual attention mechanism in his model to achieve accurate and robust crack segmentation results. Khan et al. [18] established several short connections between the encoder and decoder blocks of the UNet model to enable the architecture to obtain better pixel information flow, named Dense-UNet. Gao et al. [19] proposed an MRA-UNet, which uses a multi-scale residual module to capture fracture information at different scales on the down-sampling path, and adopted a plug-and-play dual attention module to recover features on the up-sampling path for crack segmentation.

Recently, with the increasing demand for crack measurements, portable crack measurement devices have gradually been developed. These devices typically have a compact and light-weight design. Thus, users can use these devices for crack measurements anytime and anywhere, without relying on large equipment. Portable crack measurement devices can monitor the changes of cracks in real time. They can perform continuous measurements and record crack data in order to timely grasp the development of cracks and carry out the corresponding maintenance. In addition, portable crack measurement devices usually have automated or semi-automated software, which could quickly complete crack measurement tasks, reducing manual operations and labor costs. Thus, portable crack measurement devices could improve work efficiency, saving time and energy.

However, the existing crack segmentation models in these devices have a very large size and must rely on cloud-based network computing. Therefore, in response to the issue, a light-weight and high-precision crack image segmentation model is proposed by us. The main contributions of this article are as follows:

(1) A crack image segmentation model based on a light-weight VIT module is designed by us. Due to the strong continuity of cracks, there is a certain correlation between cracks in different positions in the crack images. In our model, VIT is used to capture the relationship between different crack positions in crack images, and thereby better processing global information. More importantly, the computational complexity of the VIT module is reduced by the FSVIT. On the one hand, the fully connected layers in VIT are replaced by Depthwise Separable Convolution, and on the other hand, the Feature Space Choosing layer is used to select channels for features and reduce the number of channels for crack features.
(2) A Feature Channel Selecting Module (FCSM) is used to select channel features in the decoder. The key operation of our proposed FCSM is the Channel Sparse Choosing operation. In the Channel Sparse Choosing operation, each channel's feature corresponds to a scaling factor α and the channels with scaling factors approaching zero are pruned. Therefore, the number of channels of the original features significantly decrease after being processed by FCSM. In addition, the FCSM could suppress the influence of interfering features.
(3) The current public concrete crack segmentation datasets have too few samples, and the crack samples are very similar. Thus, these datasets may not fully reflect the crack scenarios in the real world. Therefore, this study creates a new dataset named QUCrack which contains a large number of irregular cracks in a variety of environments.

2. Materials and Methods

2.1. The Structure of the LTNet

In this article, a light-weight concrete crack segmentation model based on the Feature Sparse Choosing VIT (LTNet) is proposed by us. In Figure 1, it can be seen that the LTNet has a symmetrical structure, with the encoder on the left side and the decoder on the right side. In the encoder, the input crack image is firstly input into the stacked convolution layers for high-level feature extraction. Then these high-level features are input into the stacked Feature Sparse Choosing-based VITs (FSVITs) proposed by us. It is worth noting that the FSVIT is a light-weight VIT module. Like VIT, the FSVIT utilizes the Transformer's self-attention mechanism, which could capture global dependencies when processing features. This enables FSVIT to better learn the feature relationships between different regions in crack images. Different from VIT, the computational complexity of the FSVIT is much smaller. In the decoder, stacked convolution layers and up-sampling layers are used for non-linear feature transformation and feature size recovery. Up-sampling is commonly used to increase low-resolution images to high resolution in order to obtain more detailed information. The commonly used up-sampling methods include nearest neighbor interpolation, bilinear interpolation, and bicubic interpolation. They estimate the value of new pixels based on the values of surrounding pixels, thereby increasing the resolution of the image.

In addition, for the purpose of reducing the redundant features of the model, a Feature Channel Selected Module (FCSM) is used to reduce the number of channel features. It is noticed that Depthwise Separable Convolutions (DWConv) are used to substitute traditional convolutions for further reducing computational complexity. Detailed descriptions of these modules designed by us will be illustrated in the following sections.

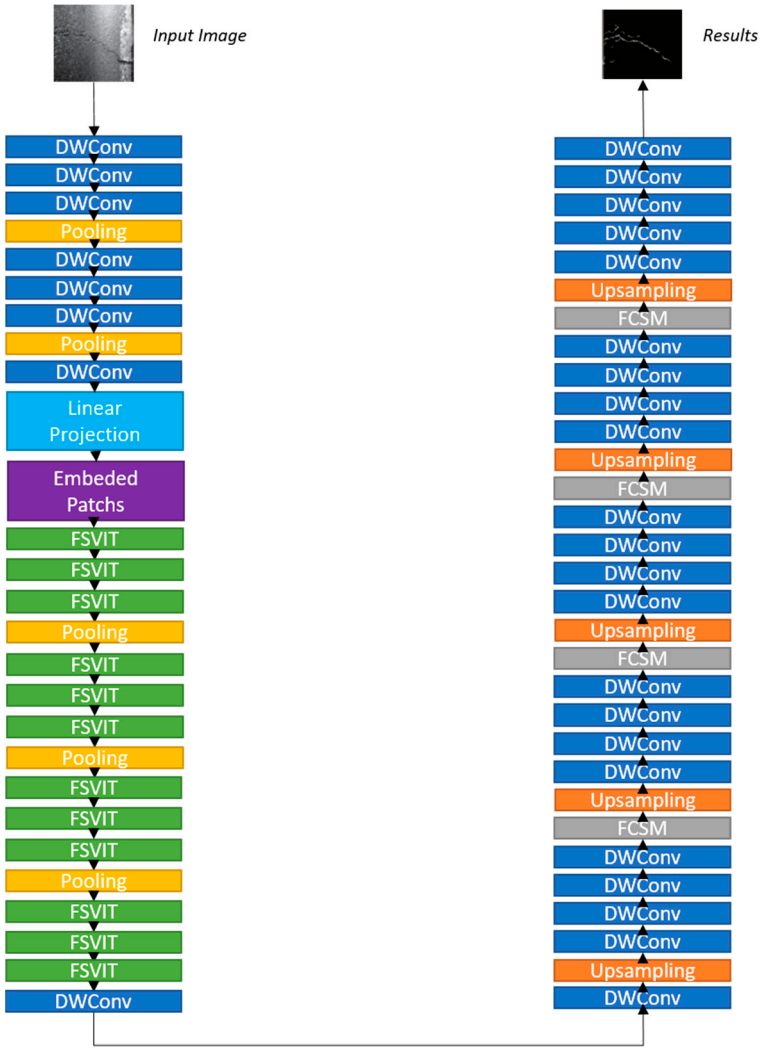

Figure 1. The structure of the LTNet.

2.2. The Feature Sparse Choosing VIT Module

VIT (Vision in Transformer) is a visual model based on the Transformer architecture. VIT treats images as a sequence and converts them into input for the Transformer model. VIT uses a self-attention mechanism to capture the relationship between different crack positions in crack images, thereby better processing global information. VIT has a strong generalization ability and transfer learning ability, which can achieve good performance in crack segmentation tasks. However, due to the complexity of the VIT model, it requires a large number of parameters and computational resources for inference. In order to reduce the computational complexity of VIT, a Feature Sparse Choosing VIT (FSVIT) is proposed by us. In our proposed FSVIT, Embedded Patches are firstly input into the Multi-Head Attention Block. The Multi-Head Attention Block maps input sequences into multiple different representations using multiple independent attention blocks, and concatenates these outputs to obtain the final attention representation of features. It is worth noting that after the Multi-Head Attention Block, a Depthwise Separable Convolution (DWConv) layer is used to replace all fully connected layers in traditional VIT to significantly reduce the computational complexity of VIT. The DWConv decomposes convolution operations into two independent steps: Depthwise Convolution and Pointwise Convolution. The DWConv only uses a large convolution kernel for each feature channel, which greatly reduces the number of parameters and computational complexity compared to fully connected layers. Then, the output features of DWConv are input into our Feature Space Choosing layer. Feature Space Choosing is another key operation for us to reduce the computational

complexity of VIT. In our Feature Space Selection layer, the global values of each feature channel are calculated through Global Average Pooling (GAP). Then, our Feature Selection layer sorts these global values in descending order, and only outputs the feature maps of the feature channels corresponding to the top 50% of the global values. In this way, by selecting feature channels, the number of output feature channels of our FSVIT would gradually decrease as the number of FSVITs increase, thus it could reduce the computational complexity of VIT when stacked FSVITs are used. The FSVIT is shown in Figure 2.

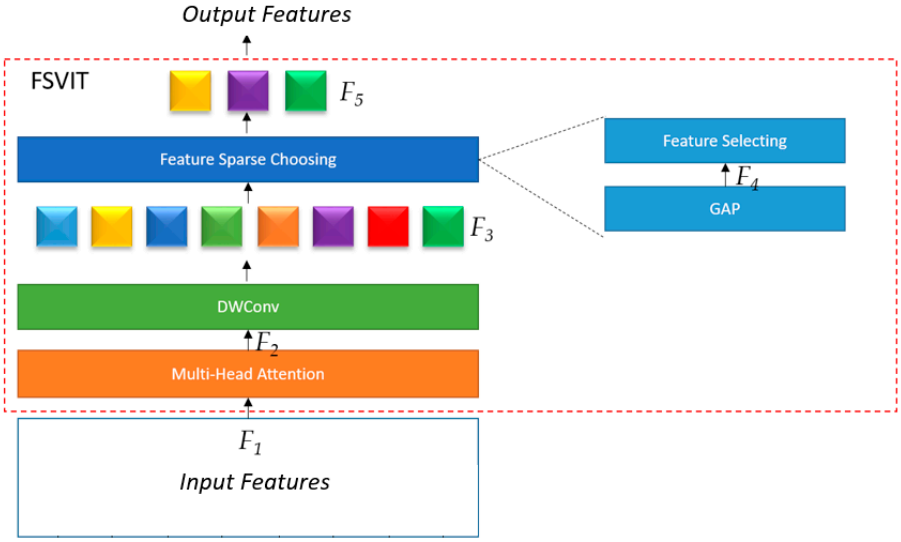

Figure 2. The structure of the Feature Sparse Choosing VIT Module (FSVIT).

The calculation of the whole process of the FSVIT is shown as follows:

$$\begin{aligned} F_2 &= MultiHeadAttention(F_1) \\ F_3 &= DWConv(F_2) \\ F_4 &= GAP(F_3) \\ F_5 &= FeatureSelecting(F_4) \end{aligned} \tag{1}$$

where $GAP(x)$ represents the Global Average Pooling operation and the $DWConv(x)$ represents the Depthwise Separable Convolution.

To illustrate the effectiveness of our FSVIT, the computational complexity of the FSVIT is compared with traditional VIT. M represents the input feature channel of the FSVIT, N represents the output feature channel of the FSVIT, where M is much smaller than N. W_Q, W_K, and W_V represent self-attention weights in Multi-Head Attention.

The parameters in the FSVIT could be calculated as follows:

$$(3 * 3 * M + 1 * N) + W_Q + W_K + W_V = 9M + N + W_Q + W_K + W_V \tag{2}$$

In traditional VIT, there usually exists two fully connected layers, with 128 neurons in each layer. However, a single Depthwise Separable Convolution is used by us to substitute the two fully connected layers. The parameters in traditional VIT could be calculated as follows:

$$\begin{aligned} &(M * 128 + N * 128 + 128 * 128) + W_Q + W_K + W_V \\ &= 128M + 128N + 16,384 + W_Q + W_K + W_V \end{aligned} \tag{3}$$

It can be seen that the computational complexity of (2) is much smaller than (3). Therefore, it can be seen that compared with the traditional VIT, the FSVIT proposed by us could reduce a large number of parameters.

2.3. The Structure of the Feature Channel Selecting Module

Feature channel selection refers to selecting which channels (or feature maps) to use in convolutional neural networks for subsequent processing and analysis. Feature channel selection has the following advantages:

(1) Reducing computational complexity: In some cases, the number of channels for inputting feature maps may be very large, resulting in higher computational complexity. By selecting specific channels, the number of channels that need to be processed can be reduced, thereby reducing computational complexity. This helps to improve the efficiency and speed of the model.
(2) Improving the generalization ability of the model: In some cases, certain channels may not have significant discriminative ability for specific tasks. By selecting channels with higher discrimination, the model's generalization ability can be improved to better adapt to new data beyond the training data.
(3) Reducing the risk of over-fitting: Excessive feature channels may increase the complexity of the model, which can easily lead to over-fitting problems. By selecting the most representative feature channels, the complexity of the model can be reduced, the risk of over-fitting can be reduced, and the generalization ability of the model can be improved.

Based on the principle of feature channel selection, a Feature Channel Selecting Module (FCSM) is proposed by us. The key operation of our proposed FCSM is the Channel Sparse Choosing operation. In the Channel Sparse Choosing operation, each channel's feature corresponds to a scaling factor α, which α is multiplied with the channel feature matrix. Next, the modified loss function L_{sum} is used to jointly train the network weights and these scaling factors α. Finally, the channels with scaling factors approaching zero are pruned, while fine-tuning the pruned network. Finally, features output from the Channel Sparse Choosing operation are shuffled and concatenated together. The L_{sum} is shown as follows:

$$L_{sum} = \sum_{(x,y)} MSE(f(x, W), y) + \lambda \sum_{\alpha} |\alpha| \tag{4}$$

where (x, y) is the training input (crack images) and training label. W is the trainable weight of the network. MSE is the mean square error loss. $f(x,y)$ is the calculation of the model, x is the input image, and W is the weight parameter. α is the channel scaling factor. λ is the balance factor. It could be seen from (4) that L1 regularization is used by us to sparsify the feature channels. L1 regularization can drive certain weights to accurately become zero, thus enabling feature weight selection. This means that the L1 regularization can automatically identify and remove features unrelated to crack targets, making the model more concise and interpretable. Through the L1 regularization, the model becomes sparser. This helps to reduce the model complexity, reducing the risk of over-fitting, and improving the model's generalization ability. The process of the FCSM is shown in Figure 3.

The calculation of the whole process of the FCSM is shown as follows:

$$\begin{aligned} B_2 &= ChannelSplit(B_1) \\ B_3 &= ChannelSparseChoosing(B_2) \\ B_4 &= Concatenate(shuffle(B_3)) \end{aligned} \tag{5}$$

where *ChannelSplit*(x) represents splitting feature x into separate channels, the *ChannelSparseChoosing*(x) represents the Channel Sparse Choosing operation and the *Concatenate*(x) represents feature concatenation. Feature concatenation is the fusion of features according to channels. The *Shuffle*(x) represents shuffling feature x randomly. Shuffling features would sort features by channel randomly.

Usually, the Channel Sparse Choosing operation only retains 30% of the features. Therefore, it could be seen that the number of channels of the original features has significantly decreased after being processed by FCSM. Therefore, the FCSM could reduce computational complexity while reducing the risk of over-fitting.

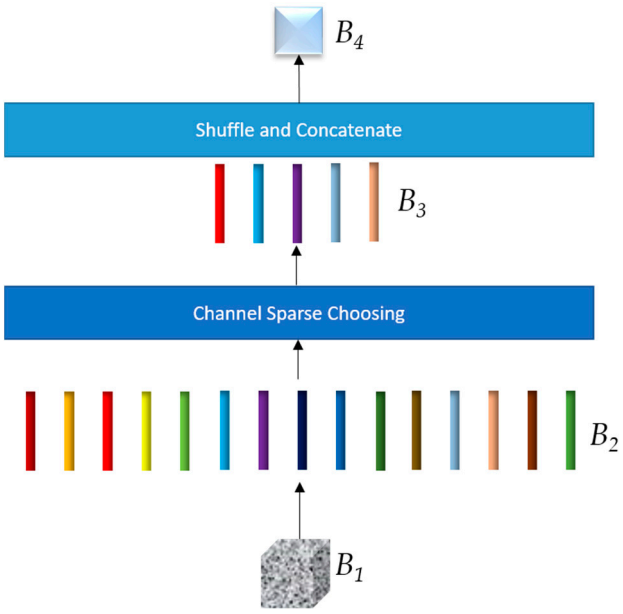

Figure 3. The structure of the Channel Selecting Module (FCSM).

2.4. Our Collected Dataset

The current public concrete crack detection dataset has too few samples, and the crack samples are very similar. These datasets may not fully reflect the crack segmentation scenarios in the real world. For example, they may not be able to capture subtle changes, noise, lighting conditions, and other factors in the real world. In addition, the existing public crack segmentation datasets lack diversity, meaning they may not be able to cover various types, shapes, and sizes of cracks. This may lead to poor performance of the model when facing unseen crack samples, as it does not have enough diverse data for learning and generalization. In response to the above situation, our own concrete crack dataset is connected and named as the QUCrack dataset. Our dataset includes concrete cracks in various environments, such as concrete house walls, concrete molds, and concrete roads. Also, our data cover multiple environmental interference scenarios such as rainy days, snowy days, shadows, strong light and nights, in order to getting closer to real crack scenarios.

Some examples of our collected crack dataset are shown in Figure 4.

(**a**)

Figure 4. *Cont.*

(**b**)

(**c**)

Figure 4. Some examples of images from our QUCrack dataset. (**a**) Crack on concrete walls. (**b**) Crack on concrete columns. (**c**) Crack on concrete roads.

3. Datasets and Experimental Setup

3.1. Datasets

Table 1 shows the splitting of the three datasets. The detailed descriptions are as follows:

Table 1. Splitting of the datasets.

Dataset	Image Size	Train	Test
The Crack500 dataset	1440 × 2560 or 2560 × 1440	1896	1124
The OAD_CRACK dataset	1920 × 1080	3500	1500
The QUCrack dataset	1920 × 1080	8000	2000

The Crack500 dataset [20]: This dataset includes 3020 images collected by Temple University, mainly capturing in campus by students. It has two kinds of size: 1440 × 2560 and 2560 × 1440.

The OAD_CRACK dataset [21]: This dataset is collected in Shenzhen, including a 5000 crack image, which is divided into four classes: linear crack, circular crack, void and background.

The QUCrack dataset: This dataset is collected by us in multiple scenarios, including concrete cracks in various environments, such as concrete house walls, concrete molds and concrete roads. Our dataset covers multiple environmental interference scenarios such as rainy days, snowy days, shadows, strong light and nights, in order to get closer to real crack scenarios.

3.2. Experimental Setup

In the experiments designed by us, all images are normalized and augmented before input into our model. In DWConv, the number of convolution filters is set to 64, 128, 256, 512, 1024, 1024, 2048, 2048 in the encoder. And in the decoder, the number of convolution filters is set to 2048, 2048, 2048, 2048, 1024, 1024, 1024, 1024, 512, 512, 512, 512, 256, 256, 256, 256, 128, 128, 128, 64, 64. Also, Stochastic Gradient Descent [21] (SGD) is used as the training policy to train our model and E-Focal Loss is used as the loss function. In addition, accuracy, recall and F1 measure are used as the evaluation criterion.

4. Results

4.1. Comparison with the State-of-the-Art Methods

To evaluate the performance of our LTNet model, a series of comparative experiments are designed by us. Accuracy, recall and F1 values are used as our basic testing standards. In addition, some mainstream crack segmentation models are adopted as our comparative algorithms. The specific introduction of these models are as follows:

ConvNet: a deep convolution-based segmentation neural network, this is the basic convolution based segmentation model.

DWTA-U-Net: a U-Net based network with discrete wavelet transformed image features for concrete crack segmentation.

CrackW-Net: a ResU-Net-based CNN for pavement crack segmentation proposed by Han.

Split-Attention Network: a channel-wise attention-based network.

DMA-Net: DeepLab With Multi-Scale attention for pavement crack segmentation proposed by Sun.

ACAU-Net: an atrous convolution and attention U-Net model for pavement crack segmentation proposed by Feng.

Cascaded Attention DenseU-Net: an attention-based network with global attention and core attention for road crack detection.

ECA-Net: a light-weight channel attention-based convolution neural network.

FU-Net: a generative adversarial network-based U-Net for road crack segmentation proposed by Gao.

Two-stage CNN: a two-stage CNN for road crack detection and segmentation proposed by Nhung.

PSNet: a Parallel Convolution-Based U-Net for Crack Detection with Self-Gated Attention Block proposed by Zhang.

PHCNet: a Pyramid Hierarchical Convolution-Based U-Net for Crack Detection with Mixed Global Attention Module and Edge Feature Extractor proposed by Zhang.

LCSNet: a light-weight Convolution Based Segmentation Method with a Separable Multi Directional Convolution Module for Concrete Crack Segmentation proposed by Zhang.

From Table 2, it could be seen that our LTNet has achieved the best performance in accuracy, recall and F1 measure, which is about 3–8% higher than the current mainstream algorithms. However, our model size is only 2 M and could be used in portable devices. These experiments fully demonstrate the effectiveness of our algorithm.

Compared with our previously proposed LCSNet, our algorithm has a 3% higher accuracy, recall and F1 measure, but the model size is the same. The reason is that the stacked VIT modules are used in our LTNet. The VIT module has the following advantages: (1) Traditional convolutional neural networks capture local features of images through local convolution operations, while the VIT module can achieve global perception through the self-attention mechanism, modeling the entire image and better capturing of global features of the image. (2) Due to the strong continuity of cracks, there is a certain correlation between cracks in different positions in crack images. The VIT module uses Transformer's self-attention mechanism to process image patches. In the self-attention mechanism, each embedded patch interacts with others, capturing the positional relationships of cracks by

calculating attention weights. In this way, the LTNet could learn correlations between cracks in different positions.

Table 2. (**a**): Accuracy comparison with the state-of-the art methods. (**b**) Recall comparison with the state-of-the art methods. (**c**) F1 measure comparison with the state-of-the art methods.

(a)

Methods	Crack500	OAD_CRACK	QUCrack	Model Size
ConvNet [22]	0.591	0.572	0.416	-
U-Net by Jenkins [23]	0.681	0.681	0.475	-
U-Net by Nguyen [24]	0.695	0.683	0.473	-
U-Net proposed by Di [25]	0.732	0.729	0.522	-
DWTA-U-Net [26]	0.77	0.754	0.567	-
CrackW-Net [27]	0.789	0.769	0.572	-
Split-Attention Network [28]	0.73	0.696	0.519	-
DMA-Net [29]	0.746	0.715	0.534	-
ACAU-Net [30]	0.792	0.774	0.575	-
Cascaded Attention DenseU-Net [31]	0.74	0.695	0.532	137 M
ECA-Net [32]	0.753	0.711	0.546	87 M
FU-Net [33]	0.795	0.769	0.582	90 M
Two-stage CNN [34]	0.79	0.765	0.585	230 M
PSNet [21]	0.812	0.762	0.612	185 M
PHCNet [35]	0.823	0.773	0.643	167 M
LCSNet [36]	0.836	0.785	0.661	2 M
LTNet	0.887	0.817	0.693	2 M

(b)

Methods	Crack500	OAD_CRACK	QUCrack
Split-Attention Network [28]	0.725	0.682	0.502
DMA-Net [29]	0.775	0.762	0.56
ACAU-Net [30]	0.776	0.753	0.551
Cascaded Attention DenseU-Net [31]	0.732	0.681	0.524
ECA-Net [32]	0.767	0.723	0.552
FU-Net [33]	0.761	0.736	0.557
Two-stage CNN [34]	0.773	0.742	0.563
PSNet [21]	0.829	0.771	0.599
PHCNet [35]	0.817	0.765	0.631
LCSNet [36]	0.828	0.773	0.652
LTNet	0.882	0.805	0.681

(c)

Methods	Crack500	OAD_CRACK	QUCrack
Split-Attention Network [28]	0.73	0.69	0.51
DMA-Net [29]	0.76	0.74	0.55
ACAU-Net [30]	0.78	0.76	0.56
Cascaded Attention DenseU-Net [31]	0.74	0.69	0.53
ECA-Net [32]	0.76	0.72	0.55
FU-Net [33]	0.78	0.75	0.57
Two-stage CNN [34]	0.78	0.75	0.57
PSNet [21]	0.82	0.77	0.61
PHCNet [35]	0.82	0.77	0.64
LCSNet [36]	0.83	0.78	0.66
LTNet	0.88	0.81	0.69

Compared with traditional attention mechanism-based models, PSNet and PHCNet could achieve better performance because crack images have strong irregularity and rich spatial features. The Pyramid Hierarchical Convolutional Module (PHCM) in PHCNet extracts multi-scale feature information from the perspective of convolutional filters, while the Parallel Convolutional Module (PCM) in PSNet extracts multi-scale feature information

from the perspective of convolutional layers. Both of these multi-scale feature information networks have a good effect on the spatial feature learning of cracks. In addition, PHCNet designed an edge feature extractor to extract the edge feature of cracks in the model to learn the edge features of cracks. Therefore, compared to ordinary attention mechanism models, PSNet and PHCNet have better performance. However, in our LTNet model, the spatial features of cracks are learned through a unified VIT architecture, thus our model does not require specially designed multi-scale feature extractors.

Compared with traditional UNet models and ConvNet models, models with attention mechanisms such as two-stage-CNN, FU-Net, ECA-Net, ACAU-Net, DMA-Net, Split-Attention Network could achieve excellent performance. The reason is that the attention mechanism can dynamically allocate attention weights based on the input contextual information, allowing the model to focus on important information and ignore irrelevant information more accurately. This operation could improve the effective utilization of features, thereby enhancing its performance.

The prediction results of our LTNet model are shown in Figure 5, and crack images are selected from Crack500 as the display images. The red boxes in the picture represent the differences in prediction results of different algorithms.

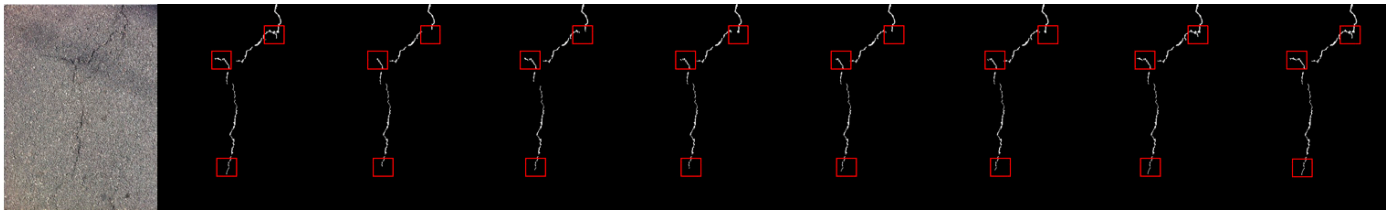

Figure 5. An example of the comparison of our proposed LTNet with the state-of-the-arts results, the example crack image is captured from the Crack500 datasets.

4.2. Effects of Using Different Numbers of FSVIT

For the purpose of evaluating the effect of using different numbers of FSVIT, an experiment is conducted.

From Figure 6, it can be seen that different numbers of FSVITs have a significant impact on the final accuracy of the model. As the number of FSVITs increases, the accuracy of the model improves rapidly. The reasons are as follows: Due to the shape of cracks or the continuity of edges, there may be a certain dependency relationship between cracks in different positions in the image. The Transformers in FSVIT use the self-attention mechanism to capture the dependency relationships between different positions in the input sequence; thus, this mechanism is particularly effective for processing crack image data. Therefore, as the number of FSVITs increases, the accuracy of the model improves rapidly. But when there are too many FSVITs, the accuracy of the model begins to slowly decline. The reason is that excessive use of the FSVIT would lead to a redundant parameter. For example, redundant FSVITs can only serve as non-linear transformations and cannot serve as feature extraction, so the parameters of these FSVITs are redundant. These redundant parameters may lead to over-fitting of the model; thus, it performs well on training data but has poor generalization ability on testing data. As a result, the model is overly sensitive to noise and subtle differences in the training data, and cannot accurately generalize to new data. Therefore, 12 FSVITs are selected for the final model.

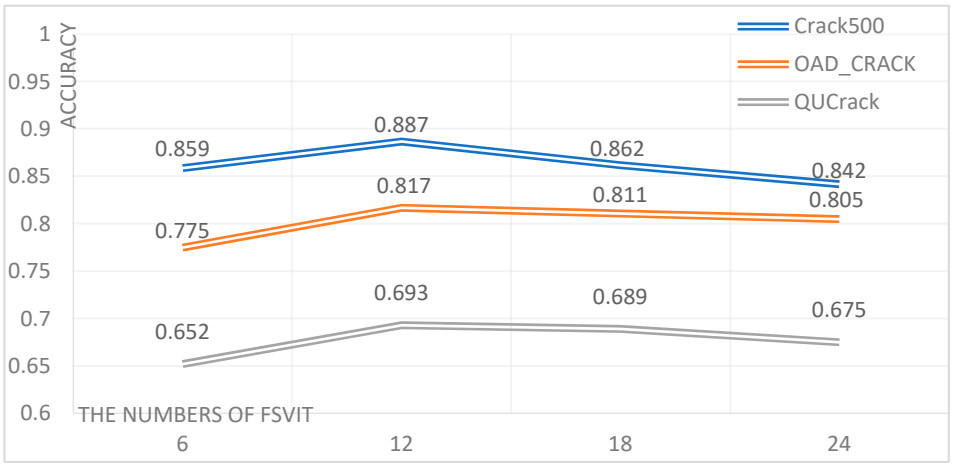

Figure 6. Accuracy comparison using different numbers of FSVIT.

4.3. Effects of Using Different Numbers of FCSM

For the purpose of evaluating the effect of using different numbers of FCSM, an experiment is conducted by us.

From Figure 7, it could be seen that using different numbers of FCSMs have a significant impact on the final accuracy of the model. As the number of FCSMs increases, the accuracy of the model improves rapidly.

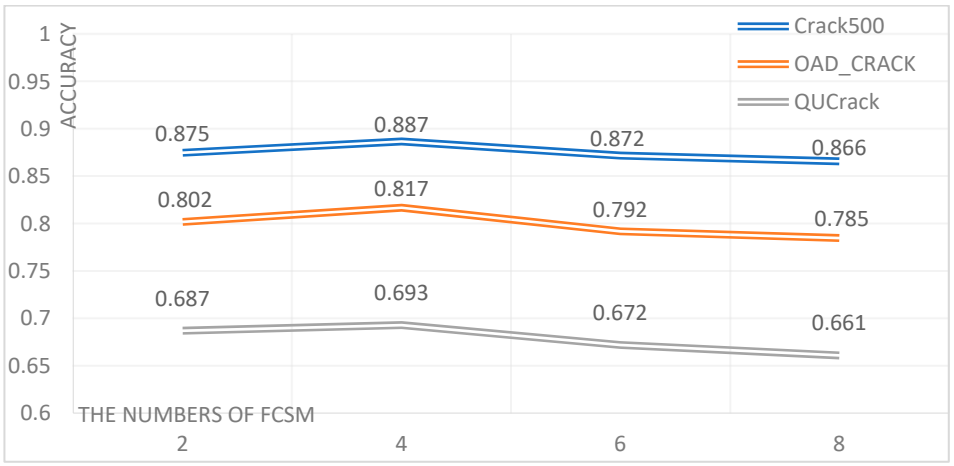

Figure 7. Accuracy comparison using different numbers of FCSM.

The reason is that by selecting appropriate feature channels, useful feature information for the task can be extracted, thereby improving the performance and generalization ability of the model. Different channels may have strong responses to different features, and selecting the appropriate channel can help the model better capture key image features. By selecting channels with strong responses to cracks, the model can more accurately locate and identify cracks, thereby improving accuracy. Additionally, selecting appropriate feature channels can reduce the interference of noise on the model. The sensitivity of different channels to noise may vary. Choosing channels that are not sensitive to noise can improve the accuracy of crack detection models.

However, excessive FCSM would lead to a decrease in the accuracy of the model. The reason is that pruning too many feature channels may lead to the loss of key crack features, such as pruning some channels that respond to non-crack areas, which may result in the model being unable to accurately distinguish between cracks and non-cracks, thereby reducing the accuracy of the crack segmentation model.

Thus, the numbers of FCSM is set to 4.

4.4. Comparison of Different Light-Weight Segmentation Models

For the purpose of evaluating the effect of our light-weight models, some other light-weight segmentation models are compared with our proposed LTNet.

From Table 3, it could be seen that our LTNet achieves the highest accuracy due to the use of a large number of light-weight VIT modules. The VIT module extracts the associated features of cracks at different positions in the crack image, which can better learn the detailed feature information of cracks. However, other light-weight models only use ordinary spatial channel attention mechanisms, which only suppress interference features in crack images and therefore have limited representation of crack features.

Table 3. Accuracy comparison with the state-of-the art light-weight models.

Methods	Crack500	OAD_CRACK	QUCrack
LiteSeg [37]	0.814	0.764	0.657
MobileNet + UNet [38]	0.786	0.724	0.591
BiSeNet v3 [39]	0.792	0.733	0.618
EGE-UNet [40]	0.803	0.747	0.621
LRNNet [41]	0.812	0.762	0.649
LCSNet [36]	0.836	0.785	0.661
LTNet	0.887	0.817	0.693

5. Conclusions

Presently, portable crack measurement devices are being rapidly developed. These devices are usually small in size, are light-weight, easy to operate and could be used for measurement in complex environments. Portable crack measurement devices could display measurement results in real time, helping users to quickly understand the situation of cracks, and take necessary measures in a timely manner. In addition, these devices are usually equipped with high-precision cameras, which can accurately measure the size, depth and shape of cracks, providing reliable measurement results. However, current portable crack measurement devices must rely on cloud computing due to the large size of crack image segmentation models. But using cloud computing requires a stable network connection between devices and cloud servers. If the network is unstable or interrupted, it may cause data transmission interruptions or delays. Moreover, cloud computing consumes a significant amount of server resources, resulting in high costs for portable crack measurement devices.

To address the above issue, we designed a light-weight crack image segmentation model, named LTNet, which could be used for portable crack measurement devices. In our model, in order to capturing the feature relationships of different crack positions in crack images, a stacked VIT module was adopted in our design of LTNet. In addition, in order to reducing the computational complexity of the VIT module, Depthwise Separable Convolution was used by us to substitute the fully connected layers. On the other hand, the Feature Space Choosing layer was adopted to select channels for crack features and reduce the number of channels of these features. Additionally, a Feature Channel Selection Module (FCSM) was designed to select channel features in the decoder of the LTNet. The FCSM could not only reduce the size of features, but also suppress the interfering features.

In addition, in order to solve the shortcomings of the existing crack image segmentation dataset, a new dataset, named QUCrack, was created by us which contains a large number of irregular cracks in a variety of environments.

A series of experiments were deigned to validate our proposed LTNet. The experimental results show that our proposed LTNet could achieve the accuracy of 0.887, 0.817 and 0.693, and achieve the recall of 0.882, 0.805 and 0.681 on three datasets, respectively, which is 3–8% higher than current mainstream algorithms. However, the model size of our LTNet is only 2 M; thus, our work perfectly achieves our goal. Although our LTNet could solve the problem of crack measuring, evaluating and repairing these cracks is an emerging field that is still in the research and development stage. For example, by using machine learning

and deep learning algorithms, data on cracks can be analyzed and predicted. By learning from a large amount of crack data, artificial intelligence can help predict the propagation and failure rate of cracks, and provide more accurate repair suggestions. Moreover, by combining artificial intelligence and robotics technology, an autonomous robot system can be developed that can automatically detect and repair cracks. These robots can detect cracks through vision and sensors, and repair them using laser, spray, or other repair methods. In addition, artificial intelligence can help design and develop intelligent materials and coatings, which can automatically perceive and repair cracks. These materials and coatings can automatically release repair agents or fillers based on the location and size of cracks, achieving self-healing functions.

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

Funding: This project was partially supported by the National Natural Science Foundation of China (No. 62071499).

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 created by ourselves can be requested from the authors.

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

References

1. Xiao, Y.; Li, J. Crack Detection Algorithm based on the Fusion of Percolation Theory and Adaptive Canny Operator. In Proceedings of the 2018 37th Chinese Control Conference (CCC), Wuhan, China, 25–27 July 2018; pp. 4295–4299.
2. Salman, M.; Mathavan, S.; Kamal, K.; Rahman, M. Pavement crack detection using the Gabor filter. In Proceedings of the 16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013), The Hague, The Netherlands, 6–9 October 2013; pp. 2039–2044.
3. Akbari, J.; Ahmadifarid, M.; Amiri, A.K. Multiple Crack Detection using Wavelet Transforms and Energy Signal Techniques. *Frat. Integrità Strutt.* **2020**, *14*, 269–280. [CrossRef]
4. Ramnivas, K.; Sachin, K.S. Crack detection near the ends of a beam using wavelet transform and high resolution beam deflection measurement. *Eur. J. Mech. A/Solids* **2021**, *88*, 104259, ISSN 0997-7538.
5. Kaul, V.; Yezzi, A.; Tsai, Y. Detecting Curves with Unknown Endpoints and Arbitrary Topology Using Minimal Paths. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 1952–1965. [CrossRef] [PubMed]
6. Hu, Y.; Zhao, C.-X. A novel LBP based methods for pavement crack detection. *J. Pattern Recognit. Res.* **2010**, *5*, 140–147. [CrossRef] [PubMed]
7. Mojidra, R.; Li, J.; Mohammadkhorasani, A.; Moreu, F.; Bennett, C.; Collins, W. Vision-based fatigue crack detection using feature tracking. *Earthq. Eng. Eng. Vib.* **2023**, *22*, 19–39. [CrossRef]
8. Aswini, E.; Divya, S.; Kardheepan, S.; Manikandan, T. Mathematical morphology and bottom-hat filtering approach for crack detection on relay surfaces. In Proceedings of the International Conference on Smart Structures and Systems Icsss'13, Chennai, India, 28–29 March 2013; pp. 108–113.
9. Nguyen, T.S.; Begot, S.; Duculty, F.; Avila, M. Free-form anisotropy: A new method for crack detection on pavement surface images. In Proceedings of the 2011 18th IEEE International Conference on Image Processing, Brussels, Belgium, 11–14 September 2011; pp. 1069–1072.
10. Oliveira, H.; Correia, P.L. Automatic road crack segmentation using entropy and image dynamic thresholding. In Proceedings of the 2009 17th European Signal Processing Conference, Glasgow, UK, 24–28 August 2009; pp. 622–626.
11. Pan, Y.; Zhang, X.; Cervone, G.; Yang, L. Detection of Asphalt Pavement Potholes and Cracks Based on the Unmanned Aerial Vehicle Multispectral Imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 3701–3712. [CrossRef]
12. Shi, Y.; Cui, L.; Qi, Z.; Meng, F.; Chen, Z. Automatic Road Crack Detection Using Random Structured Forests. *IEEE Trans. Intell. Transp. Syst.* **2016**, *17*, 3434–3445. [CrossRef]
13. Sheng, P.; Chen, L.; Tian, J. Learning-based road crack detection using gradient boost decision tree. In Proceedings of the 2018 13th IEEE Conference on Industrial Electronics and Applications (ICIEA), Wuhan, China, 31 May–2 June 2018; pp. 1228–1232.
14. Qu, Z.; Mei, J.; Liu, L.; Zhou, D.-Y. Crack Detection of Concrete Pavement With Cross-Entropy Loss Function and Improved VGG16 Network Model. *IEEE Access* **2020**, *8*, 54564–54573. [CrossRef]

15. Chaiyasarn, K.; Buatik, A.; Mohamad, H.; Zhou, M.; Kongsilp, S.; Poovarodom, N. Integrated pixel-level CNN-FCN crack detection via photogrammetric 3D texture mapping of concrete structures. *Autom. Constr.* **2022**, *140*, 104388. ISSN 0926-5805. [CrossRef]
16. Wang, S.; Pan, Y.; Chen, M.; Zhang, Y.; Wu, X. FCN-SFW: Steel Structure Crack Segmentation Using a Fully Convolutional Network and Structured Forests. *IEEE Access* **2020**, *8*, 214358–214373. [CrossRef]
17. He, M.; Lau, T.L. CrackHAM: A Novel Automatic Crack Detection Network Based on U-Net for Asphalt Pavement. *IEEE Access* **2024**, *12*, 12655–12666. [CrossRef]
18. Khan, M.A.-M.; Kee, S.-H.; Nahid, A.-A. Vision-Based Concrete-Crack Detection on Railway Sleepers Using Dense U-Net Model. *Algorithms* **2023**, *16*, 568. [CrossRef]
19. Gao, X.; Tong, B. MRA-UNet: Balancing speed and accuracy in road crack segmentation network. *Signal Image Video Process.* **2023**, *17*, 2093–2100. [CrossRef]
20. Sizyakin, R.; Voronin, V.V.; Gapon, N.; Pižurica, A. A deep learning approach to crack detection on road surfaces. In Proceedings of the Conference on Artificial Intelligence and Machine Learning in Defense Applications, Online, 21–25 September 2020. [CrossRef]
21. Zhang, X.H.; Huang, H. PSNet: Parallel-Convolution-Based U-Net for Crack Detection with Self-Gated Attention Block. *Appl. Sci.* **2023**, *13*, 9875. [CrossRef]
22. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*; IEEE: New York, NY, USA, 2017.
23. Mark, D.J.; Thomas, A.C.; Maria, I.I.; Tom, B.; Gordon, M. A deep convolutional neural network for semantic pixel-wise segmentation of road and pavement surface cracks. In Proceedings of the 2018 26th European Signal Processing Conference (EUSIPCO), Roma, Italy, 3–7 September 2018.
24. Nhung, T.H.; Nguyen, T.H.L.; Stuart, P.; Nguyen, T.T. Pavement crack detection using convolutional neural network. In *International Symposium on Information and Communication Technology*; Association for Computing Machinery: New York, NY, USA, 2018.
25. Di Benedetto, A.; Fiani, M.; Gujski, L.M. U-Net-Based CNN Architecture for Road Crack Segmentation. *Infrastructures* **2023**, *8*, 90. [CrossRef]
26. Yang, G.; Geng, P.; Ma, H.; Liu, J.; Luo, J. Dwta-unet: Concrete crack segmentation based on discrete wavelet transform and unet. In Proceedings of the 2021 Chinese Intelligent Automation Conference, Zhanjiang, China, 5–7 November 2021.
27. Han, C.; Ma, T.; Ju, H.; Huang, X.; Zhang, Y. Crackw-net: A novel pavement crack image segmentation convolutional neural network. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 22135–22144. [CrossRef]
28. Zhang, C.; Jiang, W.; Zhao, Q. Semantic segmentation of aerial imagery via split-attention networks with disentangled nonlocal and edge supervision. *Remote Sens.* **2021**, *13*, 1176. [CrossRef]
29. Sun, X.; Xie, Y.; Jiang, L.; Cao, Y.; Liu, B. Dma-net: Deeplab with multi-scale attention for pavement crack segmentation. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 18392–18403. [CrossRef]
30. Feng, J.; Li, J.; Shi, Y.; Zhao, Y.; Zhang, C. Acau-net: Atrous convolution and attention u-net model for pavement crack segmentation. In Proceedings of the 2022 International Conference on Computer Engineering and Artificial Intelligence (ICCEAI), Shijiazhuang, China, 22–24 July 2022; pp. 561–565.
31. Li, J.; Liu, Y.; Zhang, Y.; Zhang, Y. Cascaded attention denseunet (cadunet) for road extraction from very-high-resolution images. *Int. J. Geo-Inf.* **2021**, *10*, 329. [CrossRef]
32. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Hu, Q. Eca-net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
33. Gao, Z.; Peng, B.; Li, T.; Gou, C. Generative adversarial networks for road crack image segmentation. In Proceedings of the 2019 International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, 14–19 July 2019; pp. 1–8.
34. Nhung, H.T.; Nguyen, A.; Stuart, P.A.; Don, B.A.; Ha, T.L.B.; Thuy, T.N.C. Two-stage convolutional neural network for road crack detection and segmentation. *Expert Syst. Appl.* **2021**, *186*, 115718.
35. Zhang, X.; Huang, H. PHCNet: Pyramid Hierarchical-Convolution-Based U-Net for Crack Detection with Mixed Global Attention Module and Edge Feature Extractor. *Appl. Sci.* **2023**, *13*, 10263. [CrossRef]
36. Zhang, X.; Huang, H. LCSNet: Light-Weighted Convolution-Based Segmentation Method with Separable Multi-Directional Convolution Module for Concrete Crack Segmentation in Drones. *Electronics* **2024**, *13*, 1307. [CrossRef]
37. Emara, T.; Munim, H.E.A.E.; Abbas, H.M. LiteSeg: A Novel Lightweight ConvNet for Semantic Segmentation. In *2019 Digital Image Computing: Techniques and Applications (DICTA)*; IEEE: New York, NY, USA, 2020. [CrossRef]
38. Wang, B.; Li, H.S. Lane detection algorithm based on MoblieNet + UNet lightweight network. In Proceedings of the 2021 3rd International Symposium on Robotics & Intelligent Manufacturing Technology (ISRIMT), Changzhou, China, 24–26 September 2021; pp. 352–356. [CrossRef]
39. Tsai, T.H.; Tseng, Y.W. BiSeNet V3: Bilateral segmentation network with coordinate attention for real-time semantic segmentation. *Neurocomputing* **2023**, *532*, 33–42. [CrossRef]

40. Ruan, J.; Xie, M.; Gao, J.; Liu, T.; Fu, Y. Ege-unet: An efficient group enhanced unet for skin lesion segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Vancouver, BC, Canada, 8–12 October 2023; Springer Nature: Cham, Switzerland; pp. 481–490.
41. Jiang, W.; Xie, Z.; Li, Y.; Liu, C.; Lu, H. Lrnnet: A light-weighted network with efficient reduced non-local operation for real-time semantic segmentation. In *2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*; IEEE: New York, NY, USA, 2020; pp. 1–6.