symmetry

MDPI

Article

Distilling Knowledge from a Transformer-Based Crack Segmentation Model to a Light-Weighted Symmetry Model with Mixed Loss Function for Portable Crack Detection Equipment

Xiaohu Zhang and Haifeng Huang *

School of Electronic and Communication Engineering, Sun Yat-sen University, Guangzhou 510275, China; zhangxh79@mail2.sysu.edu.cn
* Correspondence: huanghaifeng@mail.sysu.edu.cn

Abstract: The detection of cracks is extremely important for maintenance of concrete structures. Deep learning-based segmentation models have achieved high accuracy in crack segmentation. However, mainstream crack segmentation models have very high computational complexity, and therefore cannot be used in portable crack detection equipment. To address this problem, a knowledge distilling structure is designed by us. In this structure, a large teacher model named TBUNet is proposed to transfer crack knowledge to a student model with symmetry structure named ULNet. In the TBUNet, stacked transformer modules are used to capture dependency relationships between different crack positions in feature maps and achieve contextual awareness. In the ULNet, only a tiny U-Net with light-weighted parameters is used to maintain very low computational complexity. In addition, a mixed loss function is designed to ensure detail and global features extracted by the teacher model are consistent with those of the student model. Our designed experiments demonstrate that the ULNet can achieve accuracies of 96.2%, 87.6%, and 75.3%, and recall of 97.1%, 88.5%, and 76.2% on the Cracktree200, CRACK500, and MICrack datasets, respectively, which is 4–6% higher than most crack segmentation models. However, the ULNet only has a model size of 1 M, which is suitable for use in portable crack detection equipment.

Keywords: crack detection; light-weighted models; crack segmentation; deep learning; symmetry model design

Citation: Zhang, X.; Huang, H. Distilling Knowledge from a Transformer-Based Crack Segmentation Model to a Light-Weighted Symmetry Model with Mixed Loss Function for Portable Crack Detection Equipment. *Symmetry* **2024**, *16*, 520. https://doi.org/10.3390/sym16050520

Academic Editor: Shangce Gao

Received: 19 March 2024
Revised: 14 April 2024
Accepted: 17 April 2024
Published: 25 April 2024

1. Introduction

Concrete is often the most common engineering material for buildings. After a period of use and exposure to the natural environment, the microstructure inside concrete may change and easily form cracks. The simplest and most commonly used method for detecting concrete cracks is to manually evaluate whether there are cracks on the surface of the concrete. However, manual observation has certain drawbacks. Artificial inspection relies on the observation and judgment of the human eye, and there are subjective issues [1]. Different people may make different judgments about cracks, which may lead to inconsistent detection results. With the development of computer vision detection, crack detection algorithms with edge detection are gradually emerging. Examples include Sobel [2], Laplace, etc. However, these algorithms require predefined thresholds, which need to be manually specified based on experience.

In the early stages of research, scholars usually used signal processing methods to detect cracks. Meghana et al. [3] proposed an image processing technology on UAVs for crack detection and assessment, in which an image stitching procedure was used to partially overlap and splice the acquired images for crack detection. Dorafshan [4] designed a model based on robot vision, using salience maps to segment cracks and automatically obtain global and local binary patterns. However, these methods rely heavily on predefined

features. Therefore, researchers have decided to seek more accurate, efficient and robust methods to identify pavement cracks.

Since the introduction of the AlexNet model in 2012, neural network models have been widely used. Liu et al. [5] designed a DeepLabv3+ based algorithm which contains Hausdorff loss and a Dual Attention Module. The key module in the DeepLabv3+ model uses ASPP, which uses dilated convolution to perform feature extraction. This operation expands the receptive field of the network while reducing the amount of calculation required. At the same time, the network uses the ASPP module to perform an attention mechanism. The ASPP can improve the model's ability to extract position-based features as well as channel-based features. Based on the above deep learning model, researchers also proposed a U-shape model, which is named UNet [6]. For example, Fan et al. [7] proposed a network for crack segmentation with an encoder-decoder based on the residual attention module, named the RAO-UNet model. Unlike other models, RAO-UNet can learn various spatial features, decomposing feature information by frequency into common features and rare features, and introduces an attention mechanism to enable the model to learn to recognize small and narrow features in crack images. However, the extended model based on UNet has some obvious shortcomings; using the interpolation method to store high-resolution crack features will limit the recovery of global information and is limited to learning edge feature information.

To solve these problems, Transformer modules have been introduced in segmentation models. The self-attention mechanism of the Transformer model makes it very powerful when processing sequence data. Thus, researchers have decided to use this model for computer vision tasks. The recent segmentation Transformer (SETR) proposed by Zheng et al. [8] uses Transformer as the CNN's encoder to extract features with generalization. Also, Yang et al. [9] designed a multi-scale network for crack detection, using the VIT module [10] to extract high-level features, and extended the receptive field of convolution, thereby achieving promising results in the segmentation of long cracks.

In recent years, portable crack detection equipment has developed rapidly [11]. These portable crack detection devices use non-contact technology, which can quickly scan and detect cracks, greatly improving the efficiency of crack detection. In addition, these devices can accurately measure the position, size, and shape of cracks, effectively avoiding human error and improving the accuracy of detection. Moreover, these devices use non-destructive detection technology, which can detect cracks without dismantling or damaging the object, avoiding potential secondary damage and safety hazards.

Additionally, due to the low cost of these portable devices, they can be easily deployed and continuously used for monitoring the changes in cracks, providing convenient and continuous crack detection for roads and buildings. However, due to the high computational complexity of crack segmentation models [12], these devices must rely on cloud computing, depending heavily on networks and cloud server resources, which cannot be deployed on a large scale. In order to solve this problem, an ultra-light-weighted crack segmentation model named ULNet is proposed by us. In addition, knowledge distilling [13] is used by us for the purpose of learning complex crack feature information, whereby a large teacher model named TBUNet is proposed to transfer crack knowledge to the ULNet (student model). The main contributions of our research are as follows:

(1) Traditional loss functions [14] used for knowledge distilling do not ensure that the detail and global features extracted by the teacher model are consistent with those of the student model. This may cause deviations in the features learned by the student model. However, crack features are very subtle and have strong global contextual information. These deviations would lead to significant changes in the final prediction generated by the student model. To solve this problem, a mixed loss function proposed by us is used to substitute the traditional loss function during the knowledge distilling process.

(2) Traditional UNet models do not consider the dependency relationships between different crack positions in feature maps, which would cause discontinuity in crack

feature extraction. Therefore, stacked transformer modules are used to capture these dependency relationships to achieve contextual awareness in our designed TBUNet.

(3) In the ULNet, only a tiny UNet with light-weighted parameters is used for maintaining very low computational complexity. In addition, compared with the traditional UNet, depth-wise separable convolutions are used to replace regular convolutions to further reduce computational complexity.

(4) The current public dataset for crack segmentation is relatively simple. To improve it, this paper presents a special crack dataset named MICrack. This dataset includes multiple angles, occlusions, and environments of cracks, meeting the needs of portable crack detection devices.

2. Methods

2.1. The Structure of the Knowledge Distilling

In this paper, for the purpose of using deep learning-based crack segmentation models in portable crack detection equipment, an ultra-light-weighted crack segmentation model (ULNet) based on knowledge distilling is presented. The main purpose of the knowledge distilling is transferring crack knowledge from the teacher model (TBUNet) to the student model with symmetry structure (ULNet). The knowledge distilling process is divided into two parts, the teacher model (TBUNet) and the student model (ULNet). TBUNet is firstly pre-trained using the crack dataset to learn crack features, and is then used for knowledge distilling. Through the process of knowledge distilling, the tiny student model (ULNet) can learn knowledge of crack features from the teacher model (TBUNet), achieving very high accuracy while maintaining low computational complexity. Details of the processes are as follows:

Firstly, in order to ensure that the details and global features of cracks extracted by the teacher model (TBUNet) are consistent with those of the student model (ULNet), a mixed loss function for supervised learning is proposed. The mixed loss function provided by us is divided into four parts: the encoder feature similarity loss L_{es}, the global feature similarity loss L_{as}, the adjustable distillation loss L_{ps} and the student loss L_{gs}. Then these four loss functions are added together to calculate the final loss during the knowledge distilling process.

Secondly, in order to get higher accuracy for crack segmentation, a transformer-based UNet with extremely large model size (TBUNet) is designed. As shown in Figure 1, in the teacher model, 12 transformer [15] encoder modules are stacked together to form a large crack feature encoder for high-level feature extraction. These transformer encoder modules use self-attention mechanisms to achieve feature encoding and feature representation, which can capture dependency relationships between different positions in feature maps and achieve contextual awareness. Note that that the TBUNet has been pre-trained using the crack dataset, and its pre-trained weights are used to teach the student model.

Finally, an ultra-light-weighted student model (ULNet) is designed by us for portable crack detection equipment. In the ULNet, a tiny UNet model with extremely light-weighted parameters is used. In addition, depth-wise separable convolutions are substituted for traditional convolutions for the purpose of further reducing model complexity.

2.2. The Mixed Loss Fuction

Traditional loss functions used for achieving knowledge distilling do not ensure the detail and global features extracted by the teacher model are consistent with those of the student model. This may cause deviations of features learned by the student model. However, the crack features are very subtle and have strong global contextual information. These deviations would lead to significant changes in the final prediction generated by the student model. To solve this problem, a mixed loss function proposed by us is used to substitute the traditional loss function during the knowledge distilling process, which is divided into four parts as shown in Figure 1.

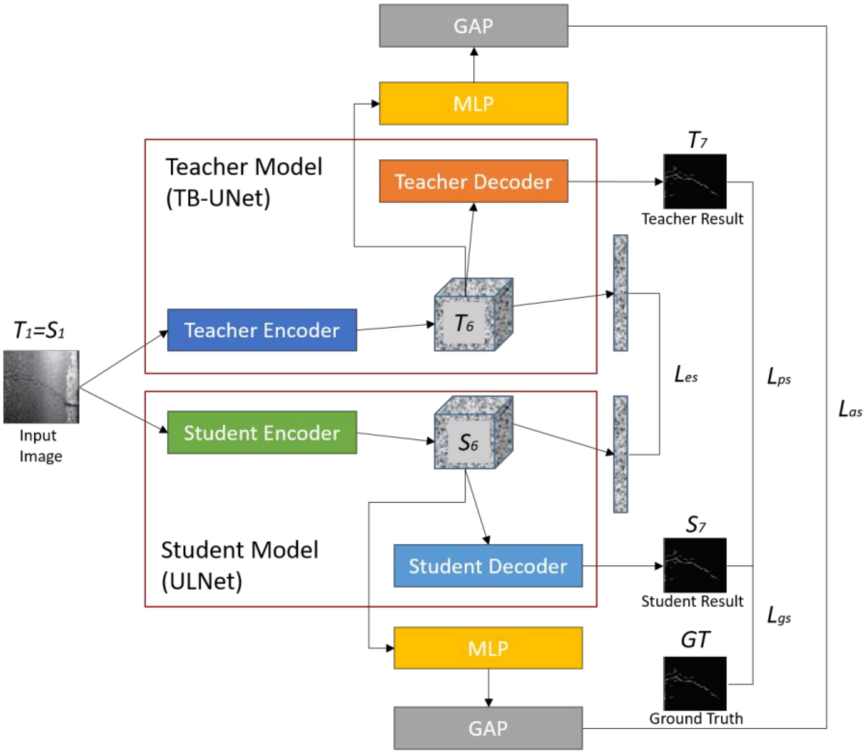

Figure 1. The knowledge distilling process with the TBUNet and the ULNet proposed by us.

(1) The encoder feature similarity loss L_{es}: The main purpose of the L_{es} is to ensure the detail features of the cracks learned by the student model are consistent with those of the teacher model. Therefore, similarity measurement is used to compare the features generated by the student model encoder with those generated by the teacher model encoder. Here, a loss function is designed by us as our encoder feature similarity loss, which is shown as follows:

$$\begin{aligned} L_{es} &= \frac{1}{N}\sum_{n=1}^{N} d^2 \\ T_{f1} &= Flatten(T_6) \\ S_{f1} &= Flatten(S_6) \\ d &= \left|\left|T_{f1} - S_{f1}\right|\right|_2 \end{aligned} \tag{1}$$

where d represents the Euclidean distance. T_6 and S_6 represent the output features of the teacher model encoder and the student model encoder respectively. N is the number of training data. *Flatten(x)* is the flatten operation.

(2) The global feature similarity loss L_{as}: The main purpose of the L_{as} is to ensure the global features of the cracks learned by the student model are consistent with those of the teacher model. First, features generated from the teacher encoder and the student encoder are input into the MLP (stacked fully connected layers) for transformation, then these two features are input into the Global Average Pooling (GAP) [16] layer for global feature extraction. Also, a loss function has been designed by us to optimize the similarity of these two global features, which is shown as follows:

$$\begin{aligned} L_{es} &= \frac{1}{N}\sum_{n=1}^{N} d^2 \\ T_{f2} &= GAP(MLP(T_6)) \\ S_{f2} &= GAP(MLP(S_6)) \\ d &= \left|\left|T_{f2} - S_{f2}\right|\right|_2 \end{aligned} \tag{2}$$

where d represents the Euclidean distance. T_6 and S_6 represent the output features of the teacher model encoder and the student model encoder respectively. N is the number of training data.

(3) The adjustable distillation loss L_{ps}: The main purpose of the L_{ps} is to maintain consistency between the prediction results of the student model and the teacher model.

In addition, the crack segmentation task usually focuses more on recall (high sensitivity), which means that cracks are predicted as much as possible without paying much attention to false detections. Therefore, a loss function is designed to mainly focus on high sensitivity, which is shown as follows:

$$L_{ps} = \frac{1}{N}\sum_{n=1}^{N}\frac{S_7 \cap T_7}{S_7 \cap T_7 + \alpha|S_7 - T_7| + \beta|T_7 - S_7|} \tag{3}$$

where T_7 is the mask predicted by the teacher model, and S_7 is the mask predicted by the student model. $|S_7\text{-}T_7|$ is false positive, $|T_7\text{-}S_7|$ is false negative, and α and β are used to control the balance between false positive and false negative. Here, α is set to 0.35 β is set to 0.65

(4) The adjustable student loss L_{gs}: The main purpose of the L_{gs} is to maintain consistency between the prediction results of the student model and the teacher model. This operation effectively brings the output of the student model closer to the real label, while preventing the knowledge learned by the teacher model from misleading the student model, which is shown as follows:

$$L_{gs} = \frac{1}{N}\sum_{n=1}^{N}\frac{S_7 \cap GT}{S_7 \cap GT + \lambda|S_7 - GT| + \theta|GT - S_7|} \tag{4}$$

where GT represents the ground truth, and S_7 is the mask predicted by the student model. $|S_7\text{-}GT|$ is false positive, $|GT\text{-}S_7|$ is false negative, and λ and θ are used to control the balance between false positive and false negative. Here, λ is set to 0.35 and θ is set to 0.65

The mixed loss function L_{all} of our method is the sum of the loss functions shown above, which can be calculated as follows:

$$L_{all} = L_{es} + L_{as} + L_{ps} + L_{gs} \tag{5}$$

2.3. The Transformer-Based UNet (TBUNet)

In order to consider the dependency relationships between different crack positions in feature maps, a transformer-based UNet (TBUNet) is used with extremely large model size.

As shown in Figure 2, the TBUNet proposed by us is an encoder-decoder-based network, with stacked transformer modules on the left and stacked convolution layers on the right. In the encoder, the crack image T_1 is firstly input into the convolution layers and pooling layers for abstract feature extraction. Then, these abstract features are split into patches with 16*16 size and flattened. After that, these flattened patches would be embedded by a Linear Projection operation to generate the embedded patches. These embedded patches are input into the stacked transformer encoders. Note that 12 transformer encoders are used to extract high-level features. These transformer encoders provide a global attention mechanism to map images into a set of sequence data. Then, a multi-head attention mechanism is used to establish associations between these embedded patches, and to learn feature representation on a global scale. Thus, these transformer encoders can effectively capture global information in images without relying on convolution operations. Finally, feature T_6 generated from the stacked transformer encoders is input into the decoder. In the decoder, stacked convolution layers are used for feature transformation and feature size restoration, generating the final result T_7. It can also be seen that several features (T_2, T_3, T_4, T_5) generated from the encoder are concatenated with features generated from the decoder.

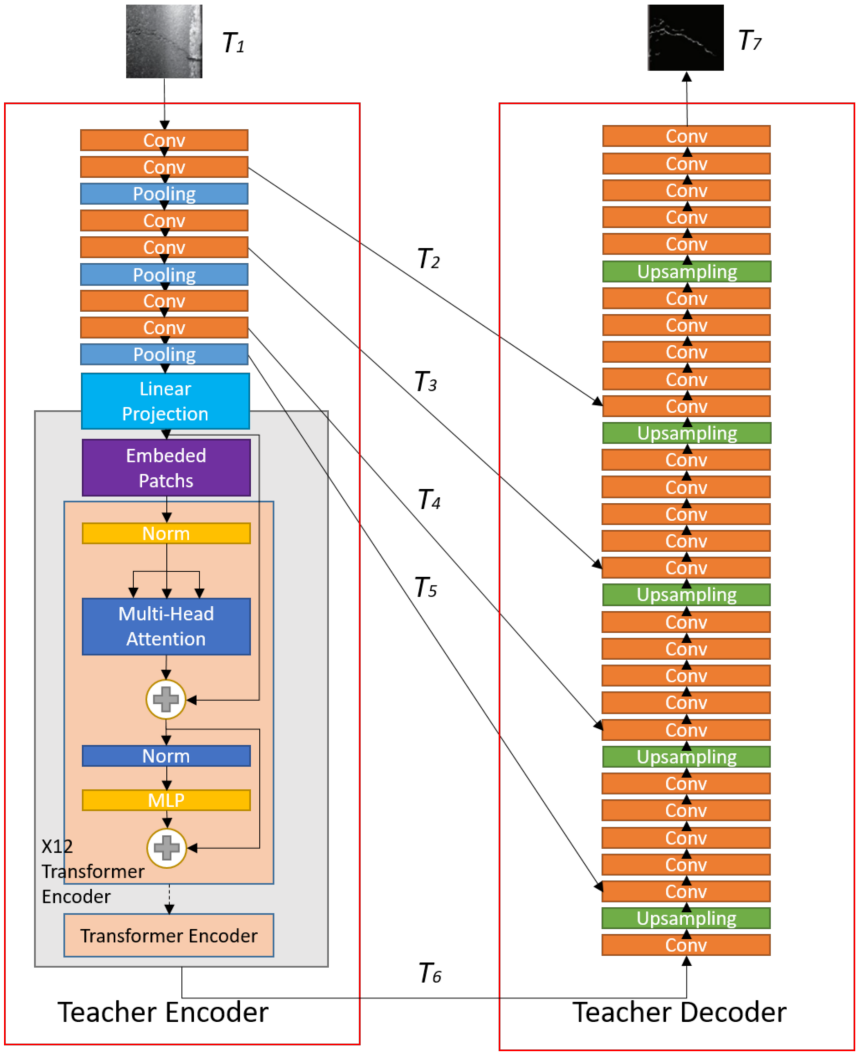

Figure 2. The Transformer-based UNet (TBUNet).

2.4. The Ultra-Light-Weighted Model (ULNet)

In order to apply crack segmentation models in portable crack detection equipment, an ultra-light-weighted model with symmetry structure (ULNet) has been designed by us. As shown in Figure 3, the ULNet proposed by us is also an encoder-decoder-based network. It can be seen that some convolution layers are used in the encoder for high-level feature extraction. In the decoder, it can be seen that some up-sampling layers are used for feature transformation and feature size restoration. For the purpose of further reducing parameters, depth-wise separable convolutions (DWConvs) [17] are used. The DWConvs firstly perform spatial convolution, using only one spatial filter for each channel of the input feature map, and then concatenate the outputs together. Subsequently, many $1 * 1$ channel filters are used for channel convolutions to generate channel feature maps. In addition, several features (S_2, S_3, S_4, S_5) generated from the encoder are concatenated with features generated from the decoder.

2.5. Dataset Collection

Current public datasets for crack segmentation are relatively simple due to their single collection method, usually taken with a camera from a plan view. However, portable crack detection devices typically capture crack photos from multiple angles. In addition, cracks in these public datasets are completely free from obscuration by external objects. However, actual cracks are often obscured and exhibit interference from vehicles, shadows, and pedestrians. Therefore, these current public datasets cannot be applied in portable

crack detection devices. In response to the above issues, a handheld camera is employed to capture crack images from multiple views, occlusions, and environments, meeting the needs of portable crack detection devices. Our dataset is mainly collected from urban sidewalks, rural cement roads, and internal roads in industrial parks, with a total of 5000 images, and is named the MICrack dataset. Here, some examples of crack images in our dataset are shown, as Figure 4.

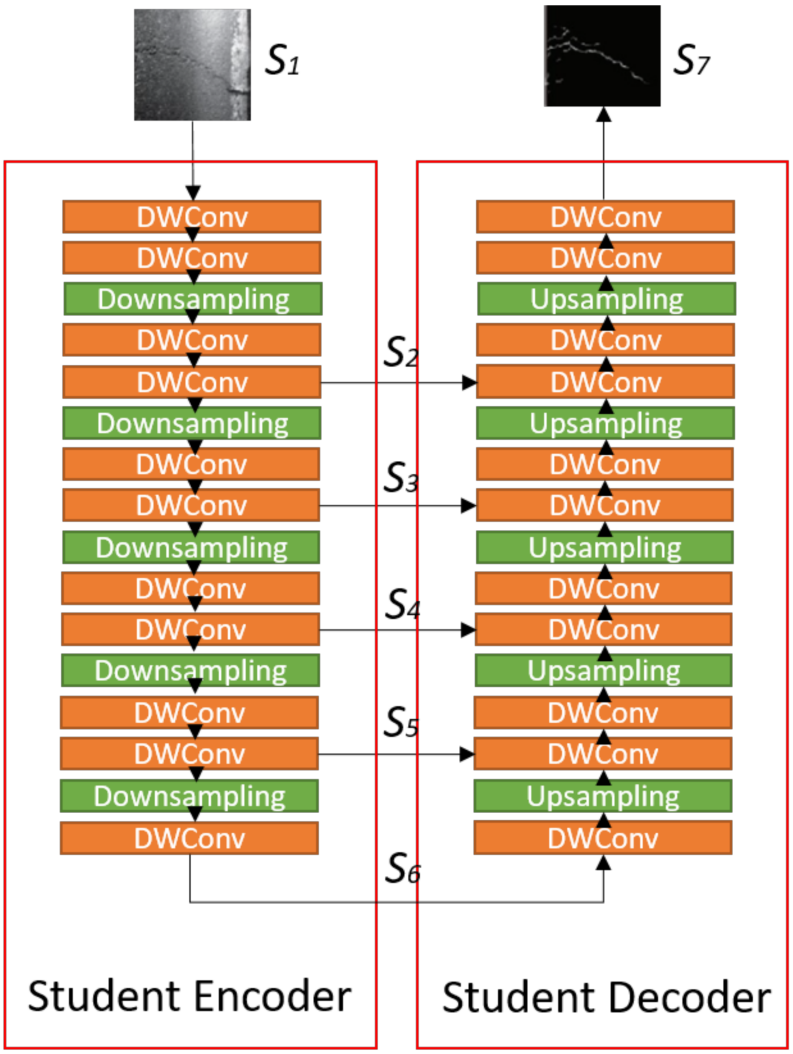

Figure 3. The ultra-light-weighted model (ULNet).

(**a**)

Figure 4. *Cont.*

(**b**)

(**c**)

Figure 4. Crack images in the MICrack dataset. (**a**) Crack with interference from road lines. (**b**) Bent cracks. (**c**) Crack which has curled and warped.

3. Datasets and Experimental Setup

3.1. Datasets

In this paper, three crack datasets are used in our experiments. All training sets and testing sets are defined by the authors who published the dataset. The details of these datasets are as follows:

Cracktree200 [18]: The collection of this crack dataset was mainly carried out by taking photos of roads from airplanes, and then filtering out images with cracks. After that, the dataset was created by manually annotating these crack images. The characteristic of this dataset is that the cracks are all gray concrete with a relatively homogenous color.

Crack500 [19]: This crack dataset was mainly collected at Temple University, where students manually captured images using cameras and then filtered and annotated the images. Due to the different resolutions of the cameras, this dataset has two sizes of crack images.

MICrack: This crack dataset is made by the authors of this paper. Crack images were collected from urban sidewalks, rural cement roads, and internal roads in industrial parks, with a total of 5000 images. The interference within this dataset is very complex.

The discriptions of our datasets are shown in Table 1.

Table 1. Descriptions of the datasets.

Crack Dataset	Resolution of Images	Train	Test	Train:Test
Cracktree200	800 width and 600 height	165	41	8:2
Crack500	2560 width and 1440 height/1440 width and 2560 height	1896	1124	6:4
MICrack	1920 width and 1080 height	4000	1000	4:1

3.2. Experimental Setup in Our Methods

In the ULNet, the number of $1 * 1$ channel filters of the DWConv is set to 256, and the size of the spatial filter in DWConv is set to $3 * 3$. Also, ReLU is used as the activation function in ULNet. In the TBUNet proposed by us, the number of convolution filters in the encoder is set to 32, 64, 128, 1024, 2048, 2048, for each convolution layer, respectively. In the decoder, the number of convolution filters is set to 2048, 2048, 2048, 2048, 2048, 1024, 1024, 1024, 1024, 1024, 128, 128, 128, 128, 128, 64, 64, 64, 64, 64, 32, 32, 32, 32, 32, respectively. The size of all filters in the convolution layers of the TBUNet is set to $3 * 3$. Note that for the purpose of retaining negative crack features, Swish [20] activation functions are used in the TBUNet. Finally, a mixed loss function is used during the training process of the knowledge distilling. As for training, SGD [21] is used as the training policy.

3.3. Evaluation Criteria

Accuracy, recall, and F1 measure are used as the evaluation criteria for our ULNet. The description of these evaluation criteria are as follows:

(1) Accuracy

Accuracy refers to the ratio of the number of correctly predicted samples by the classifier to the total number of samples. It measures the classifier's ability to predict correctly. The formula can be expressed as follows:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

(2) Recall

Recall refers to the ratio of the number of correctly predicted positive samples by the classifier to the actual number of positive samples. It measures the classifier's ability to recognize positive examples. The formula can be expressed as follows:

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

(3) F1-measure

F1-measure is a comprehensive evaluation indicator of accuracy and recall. It is the harmonic average of accuracy and recall, used to measure the overall performance of the classifier. The formula can be expressed as follows:

$$F1 = \frac{2 * Acc * Recall}{Acc + Recall} \tag{8}$$

4. Results

4.1. Comparison with the Main Stream Crack Segmentation Models

In order to evaluate the performance of our proposed ULNet, several experiments were designed. For comparison, segmentation accuracy, recall, and F1 measure are used as the main indicators in our experiments. In addition, several mainstream crack segmentation methods proposed by other researchers are used as our baseline.

As shown in Table 2, it can be seen that our ULNet exhibits superior performance compared with these baselines. However, the computational complexity of our model is extremely low. The reasons are as follows:

- Knowledge distilling is used to transfer the crack knowledge learned by the teacher model to the student model. Here, our teacher model is a complex and accurate model, while the student model is a simplified light-weighted model. Knowledge distilling can effectively supplement the ability of student models to extract high-level features, reducing model complexity while maintaining good performance.
- In addition, knowledge distillation loss function is improved by using multiple loss functions to supervise the knowledge transfer of high-level detail features of cracks,

the knowledge transfer of the global features of cracks, the knowledge transfer of the prediction results, and the knowledge transfer of the Ground Truth. These loss functions are added together for overall training, ensuring the detail features and global features of cracks remain consistent between the teacher model and the student model during the knowledge distillation process.

- In the teacher model, stacked transformer layers are designed to learn the dependency relationships between different crack positions in feature maps. This design ensures the continuity of crack feature extraction, which improves final accuracy.

Table 2. (**a**): Comparison with mainstream crack segmentation models (Accuracy). (**b**): Comparison with mainstream crack segmentation models (Recall). (**c**): Comparison with mainstream crack segmentation models (F1-measure).

(**a**)

Methods	**Cracktree200**	**Crack500**	**MICrack**	**Model Size**
ConvNet [19]	0.471	0.591	0.392	-
U-Net by Jenkins [20]	0.75	0.681	0.519	-
U-Net by Nguyen [21]	0.763	0.695	0.531	-
U-Net proposed by Di [22]	0.791	0.732	0.546	-
DWTA-U-Net [23]	0.90	0.77	0.671	-
CrackW-Net [24]	0.855	0.789	0.632	-
Split-Attention Network [25]	0.851	0.73	0.563	-
DMA-Net [26]	0.793	0.746	0.58	-
ACAU-Net [27]	0.861	0.792	0.62	-
Cascaded Attention DenseU-Net [28]	0.863	0.74	0.598	137 M
ECA-Net [29]	0.885	0.753	0.617	87 M
FU-Net [30]	0.89	0.795	0.661	90 M
Two-stage-CNN [31]	0.892	0.79	0.664	230 M
PSNet [32]	0.926	0.812	0.681	185 M
PHCNet [33]	0.929	0.823	0.702	167 M
ULNet	0.962	0.876	0.753	1 M

(**b**)

Methods	**Cracktree200**	**Crack500**	**MICrack**
Split-Attention Network	0.857	0.725	0.557
DMA-Net	0.823	0.775	0.614
ACAU-Net	0.854	0.776	0.61
Cascaded Attention DenseU-Net	0.853	0.732	0.58
ECA-Net	0.891	0.767	0.628
FU-Net	0.864	0.761	0.643
Two-stage-CNN	0.851	0.773	0.652
PSNet	0.932	0.829	0.693
PHCNet	0.914	0.817	0.698
ULNet	0.971	0.885	0.762

(**c**)

Methods	**Cracktree200**	**Crack500**	**MICrack**
Split-Attention Network	0.85	0.73	0.56
DMA-Net	0.81	0.76	0.60
ACAU-Net	0.86	0.78	0.61
Cascaded Attention DenseU-Net	0.86	0.74	0.59
ECA-Net	0.89	0.76	0.62
FU-Net	0.88	0.78	0.65
Two-stage-CNN	0.87	0.78	0.66
PSNet	0.93	0.82	0.69
PHCNet	0.92	0.82	0.70
ULNet	0.97	0.88	0.76

Comparing these mainstream attention-based segmentation models with traditional segmentation models, it can be seen that these attention-based segmentation models could exhibit superior performance. The reason is that the attention mechanism allows the model to focus on certain specific parts of the input image. This is achieved by introducing relevant structures into the CNN model, which can calculate the importance of each pixel and adjust the model processing based on its importance, thereby helping the model pay more attention to important information within the crack image and reducing the impact of background interference on the model.

Here, a comparison of the prediction results is shown in Figure 5. The differences in the red box indicate that our model performs well in predicting crack details, and the crack segmentation results predicted by our model are very consistent, indicating the effectiveness of our model.

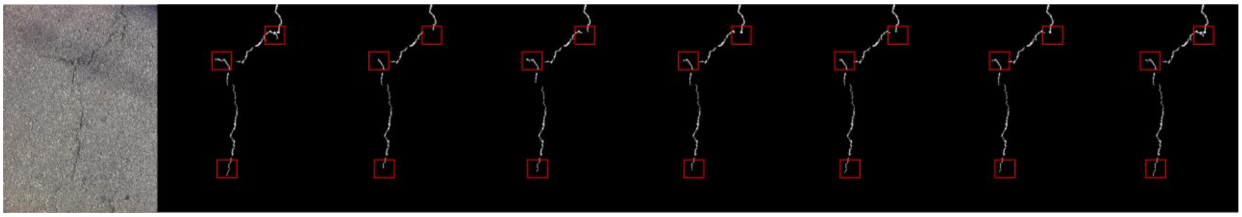

Figure 5. Comparison of visualization results of different algorithms.

4.2. Effects of Using Different Numbers of 1 ∗ 1 Filters in DWConv in ULNet

In order to explore smaller semantic segmentation models and balance the relationship between accuracy and model size, an experiment was designed to test the accuracy of the ULNet by adjusting the number of $1 * 1$ channel filters in the DWConv in ULNet.

From Figure 6, it can be seen that adjusting the number of channel filters does indeed affect the final accuracy of the model. As the number of channel filters increases, the accuracy of the model begins to rapidly improve. However, as the number of channel filters continues to increase, the accuracy of the model improves more slowly. Having too many channel filters would increase the computational complexity of the model, so the number of channel filters in DWConv is set to 256.

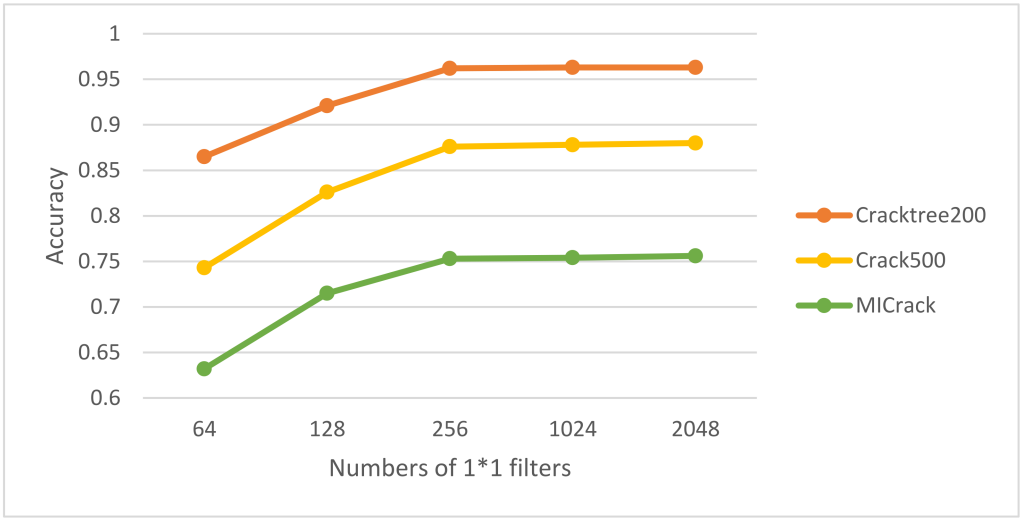

Figure 6. Effects of using different numbers of 1*1 filters in DWConv in ULNet.

4.3. Comparison of the Mainstream Light-Weighted Models

To demonstrate the powerful performance of our LCSNet, we compared our model with current mainstream lightweight models

From Table 3, it can be seen that our proposed ULNet is better than the current mainstream light-weighted semantic segmentation models, because cracks have long correlated positional features which contain temporal contextual information and require special structures to be designed for their extraction. However, the structure for extracting temporal contextual information usually has relatively high computational complexity and is difficult to use in light-weighted models. Therefore, these light-weighted models typically do not include these structures. However, our model adopts a knowledge distillation approach, using a stacked transformer structure for extracting temporal contextual feature information of position features, and using a specially designed loss function to transfer this temporal contextual feature information to our light-weighted model. Therefore, our model can achieve much higher accuracy than these mainstream light-weighted semantic segmentation models.

Table 3. Comparison of mainstream light-weighted models (Accuracy).

Methods	Cracktree200	Crack500	MICrack
LiteSeg [34]	0.925	0.814	0.703
MobileNet+UNet [35]	0.892	0.786	0.665
BiSeNet v3 [36]	0.919	0.792	0.672
EGE-UNet [37]	0.928	0.803	0.687
LRNNet [38]	0.937	0.812	0.715
ULNet	0.962	0.876	0.753

4.4. Effects of Using Different Parts in Our Designed Mixed Loss Function

To verify the effectiveness of our proposed mixed loss function, a comparative experiment was designed.

From Table 4, it can be seen that a loss function with four parts would achieve the best results. The reason is that crack features are very subtle and have strong global contextual information. Therefore, in the process of knowledge distillation, it is important to ensure the consistency of detail and global features between the teacher model and the student model. It is important to design a special loss function to perform supervised transferal of these features. For supervised transferal of detail feature information, the similarity between output features of the encoder layer of the teacher model and the student model is compared. For supervised transferal of global feature information, global average pooling is used on the output features generated from the encoder layer of the teacher model and the student model, and then the similarity of the features generated by the global average pooling are compared.

Table 4. Accuracy comparison of using different parts in our designed mixed loss function.

Loss Function	Cracktree200	Crack500	MICrack
$L_{all} = L_{ps} + L_{gs}$	0.935	0.839	0.723
$L_{all} = L_{es} + L_{ps} + L_{gs}$	0.947	0.858	0.736
$L_{all} = L_{as} + L_{ps} + L_{gs}$	0.951	0.863	0.742
$L_{all} = L_{es} + L_{as} + L_{ps} + L_{gs}$	0.962	0.876	0.753

5. Conclusions

Due to the high computational complexity of crack segmentation models, portable crack detection equipment must rely on cloud computing, depending heavily on networks and cloud server resources. To solve this, a knowledge distilling structure is proposed by us. In this structure, a large teacher model named TBUNet is proposed to transfer crack knowledge to a tiny student model named ULNet. In addition, a mixed loss function is proposed by us to be applied during the knowledge distilling process. The purpose of this loss function is to ensure that the detail and global features extracted by the teacher model are consistent with those of the student model. Also, stacked transformer modules are used

in the TBUNet. The purpose of these modules is to capture the dependency relationships between different crack positions in feature maps. As for the ULNet, a tiny U-Net proposed by us with light-weighted parameters is used for final deployment. Meanwhile, for the purpose of meeting the needs of portable crack detection devices, a new dataset was created by us which is named MICrack, containing multiple angles and occlusions. Experimental results show that the ULNet used for final deployment could achieve accuracies of 96.2%, 87.6%, and 75.3% on the Cracktree200, CRACK500, and MICrack datasets, respectively, which is 4–6% better than mainstream models. However, the ULNet's model size is 1 M, which is suitable for use in portable crack detection equipment. Therefore, our method perfectly solves the stated problem. However, more research is needed on evaluating and repairing cracks, therefore our future work will focus on this direction.

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.; visualization, X.Z.; supervision, X.Z.; project administration, X.Z. and H.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 62071499.

Data Availability Statement: All datasets may be requested by contacting the corresponding author.

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

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

References

1. Wu, C.; Sun, K.; Xu, Y.; Zhang, S.; Huang, X.; Zeng, S. Concrete crack detection method based on optical fiber sensing network and microbending principle. *Saf. Sci.* **2019**, *117*, 299–304. [CrossRef]
2. Bradski, G.; Daebler, A. *Learning OpenCV: Computer Vision with OpenCV Library*; University of Arizona: Tucson, AZ, USA, 2008.
3. Meghana, R.K.; Apoorva, S.; Mohana; Chitkara, Y. Inspection, Identification and Repair Monitoring of Cracked Concrete Structure—An Application of Image Processing. In Proceedings of the 2018 3rd International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 15–16 October 2018; pp. 1151–1154.
4. Dorafshan, S.; Maguire, M.; Thomas, R.J. *SDNET2018: A Concrete Crack Image Dataset for Machine Learning Applications*; Utah State University: Logan, UT, USA, 2018. [CrossRef]
5. Liu, J. Road Crack Detection Using HDD LOSS and Dual Attention Module with DeepLabv3+. In Proceedings of the 2023 3rd International Conference on Digital Society and Intelligent Systems (DSInS), Chengdu, China, 10–12 November 2023; pp. 148–152.
6. Zhou, S.; Wang, Q.; Wu, H.; Wang, Q.; Meng, Y.; Shen, T. ASSA-UNet: An Efficient UNet-Based Network for Chip Internal Defect Detection. In Proceedings of the 2023 11th International Conference on Information Systems and Computing Technology (ISCTech), Qingdao, China, 30 July–1 August 2023. [CrossRef]
7. Fan, L.; Zhao, H.; Li, Y.; Li, S.; Zhou, R.; Chu, W. RAO-UNet: A residual attention and octave UNet for road crack detection via balance loss. *IET Intell. Transp. Syst.* **2022**, *16*, 332–343. [CrossRef]
8. Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P.H.; et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 6877–6886.
9. Yang, Y.; Niu, Z.; Su, L.; Xu, W.; Wang, Y. Multi-scale feature fusion for pavement crack detection based on Transformer. *Math. Biosci. Eng.* **2023**, *20*, 14920–14937. [CrossRef] [PubMed]
10. Aso, R.; Shiota, S.; Kiya, H. Enhanced Security with Encrypted Vision Transformer in Federated Learning. In Proceedings of the 2023 IEEE 12th Global Conference on Consumer Electronics (GCCE), Nara, Japan, 10–13 October 2023. [CrossRef]
11. Cao, T.; Hu, J.; Liu, S. Enhanced Edge Detection for 3D Crack Segmentation and Depth Measurement with Laser Data. *Int. J. Pattern Recognit. Artif. Intell.* **2022**, *36*, 2255006. [CrossRef]
12. Zhang, E.; Shao, L.; Wang, Y. Unifying transformer and convolution for dam crack detection. *Autom. Constr.* **2023**, *147*, 104712. [CrossRef]
13. Chen, Z.; Cai, C.; Zheng, T.; Luo, J.; Xiong, J.; Wang, X. RF-Based Human Activity Recognition Using Signal Adapted Convolutional Neural Network. *IEEE Trans. Mob. Comput.* **2021**, *22*, 487–499. [CrossRef]
14. Kang, J.; Wang, Z.; Zhu, R.; Xia, J.; Sun, X.; Fernandez-Beltran, R.; Plaza, A. DisOptNet: Distilling Semantic Knowledge From Optical Images for Weather-Independent Building Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4706315. [CrossRef]

15. 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]
16. Maurya, A.; Chand, S. A global context and pyramidal scale guided convolutional neural network for pavement crack detection. *Int. J. Pavement Eng.* **2023**, *24*, 2180638. [CrossRef]
17. Mercioni, M.A.; Holban, S. P-Swish: Activation Function with Learnable Parameters Based on Swish Activation Function in Deep Learning. In Proceedings of the 2020 International Symposium on Electronics and Telecommunications (ISETC), Timișoara, Romania, 5–6 November 2020. [CrossRef]
18. Qin, C.; Li, B.; Han, B. Fast brain tumor detection using adaptive stochastic gradient descent on shared-memory parallel environment. *Eng. Appl. Artif. Intell.* **2023**, *120*, 105816. [CrossRef]
19. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *25*, 15263.
20. Jenkins, M.D.; Carr, T.A.; Iglesias, M.I.; Buggy, T.; Morison, G. 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), Rome, Italy, 3–7 September 2018.
21. Nguyen, N.T.H.; Le, T.H.; Perry, S.; Nguyen, T.T. Pavement crack detection using convolutional neural network. In Proceedings of the International Symposium on Information and Communication Technology, Da Nang, Vietnam, 6–7 December 2018.
22. Di Benedetto, A.; Fiani, M.; Gujski, L.M. U-Net-Based CNN Architecture for Road Crack Segmentation. *Infrastructures* **2023**, *8*, 90. [CrossRef]
23. 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 2021 Chinese Intelligent Automation Conference*; Deng, Z., Ed.; Lecture Notes in Electrical Engineering; Springer: Singapore, 2022; Volume 801.
24. Han, C.; Ma, T.; Huyan, J.; 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]
25. 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]
26. 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]
27. Jun, F.; 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.
28. 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]
29. 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.
30. 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.
31. Nhung Hong Thi Nguyen, A.; Stuart Perry, A.; Don Bone, A.; Ha Thanh Le, B.; Thuy Thi Nguyen, C. Two-stage convolutional neural network for road crack detection and segmentation. *Expert Syst. Appl.* **2021**, *186*, 115718. [CrossRef]
32. Zhang, X.; Huang, H. PSNet: Parallel-Convolution-Based U-Net for Crack Detection with Self-Gated Attention Block. *Appl. Sci.* **2023**, *13*, 9875. [CrossRef]
33. 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]
34. Emara, T.; Munim HE, A.E.; Abbas, H.M. LiteSeg: A Novel Lightweight ConvNet for Semantic Segmentation. In Proceedings of the 2019 Digital Image Computing: Techniques and Applications (DICTA), Perth, Australia, 2–4 December 2019; IEEE: Piscataway, NJ, USA, 2020. [CrossRef]
35. 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]
36. 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]
37. Ruan, J.; Xie, M.; Gao, J.; Liu, T.; Fu, Y. Ege-unet: An efficient group enhanced unet for skin lesion segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Cham, Switzerland, 2023; pp. 481–490.
38. 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 Proceedings of the 2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), London, UK, 6–10 July 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–6.