STUCNET – SWIN TRANSFORMER-V2 UNET FOR CRACK SEGMENTATION NETWORK

Hai-Hong Phan[1,*], *Le Hoang Tung Nguyen*[1]

DOI: 10.56651/lqdtu.jst.v12.n1.657.ict

Abstract

Automatic crack detection on road surfaces is an important task for supporting the quality control of road infrastructure in transportation. Various methods have been proposed for crack segmentation, but their accuracy is still limited. To improve the effectiveness of crack detection, we propose the Swin Transformer-V2 UNET for Crack Segmentation Network model (STUCNet) for crack recognition. The proposed model combines the advantages of the Swin Transformer-V2 into the encoding module of the UNET-based architecture to enhance the quality of semantic image segmentation. Specifically, the model integrates Swin Transformer-V2 with shifted windows as the encoder to extract contextual features for crack segmentation. The symmetric decoder is based on a convolutional neural network with attention designed to perform upsampling operations to restore the spatial resolution of the feature map. We evaluate the STUCNet model on a large dataset containing cracks collected in different contexts. Compared to current advanced models, the proposed method achieves state-of-the-art (SOTA) results for crack segmentation.

Index terms

Road crack detection, crack segmentation, Transformer, Attention, UNET.

1. Introduction

Detecting road cracks is the process of inspecting and identifying cracks on the road surface to assess and develop maintenance plans to ensure safety for road users. Detecting cracks on the road can be done manually using human eyes or automatically using computer vision. Manual inspection and evaluation are labor-intensive and time-consuming. They require the inspector to be an expert in the field of road construction. Automatic crack detection has been researched for many decades with many challenges such as noisy images, and small diverse cracks. This article proposes a novel approach based on the Vision Transformer approach to automatically detect and segment road cracks, helping to assess road conditions more accurately.

The methods of crack segmentation and detection for road images can be divided into two main groups as follows: traditional crack recognition methods and advanced crack

[1] Institute of Information and Communication Technology, Le Quy Don Technical University

[*] Corresponding author, email: hongpth@lqdtu.edu.vn

recognition methods. Traditional methods mainly rely on image processing techniques, while advanced methods automatically segment cracks using deep learning networks.

Over the past few years, there has been a continuous development of computer vision-based crack detection algorithms. Many algorithms such as threshold segmentation [1], morphology [2], and filter-based algorithms [3] have been employed for crack detection tasks. While these methods can attain quite high detection accuracy after fine-tuning the parameters, they are only suitable for images captured in a specific environment. The detection accuracy can be lowered by illumination and shooting distance, making them unsuitable for practical engineering requirements.

To overcome the limitations of traditional computer vision algorithms, Convolutional Neural Networks (CNNs) are being utilized for various vision tasks. CNNs were initially proposed by LeCun et al. [4], and can generate hierarchical features in an image, thus improving the understanding of objects within the image. CNNs have been widely applied in various areas such as image classification [5], [6], object detection [7], and semantic segmentation [8]–[10]. Among CNN variants, the UNET [11], a typical encoder-decoder-based network, has exhibited exceptional segmentation potential. In this network, the encoder extracts features by continuously down-sampling the input image, and then the decoder progressively utilizes the output of the features from the encoder through skip connections to up-sample the image. This allows the network to obtain features of varying granularity, resulting in better segmentation. With the popularity of UNET, several novel models such as SegNet [12], CrackNet [13], DeepCrack [14], and SCCDNet [15] have been proposed, specifically designed for crack image segmentation. These models have achieved remarkable performance.

Convolutional kernels have inherent inductive biases that limit their ability to consider the entire image, resulting in a loss of global context and the inability to establish long-range dependencies. Although stacking convolution layers and downsampling can help to expand the receptive field and improve local interaction, this approach is not optimal because it increases the model's complexity and makes it more prone to overfitting.

The Transformer [16], a novel architecture initially designed for sequence-to-sequence modeling in natural language processing (NLP) tasks, has recently sparked considerable discussion in the computer vision (CV) community. The Transformer can potentially revolutionize most NLP tasks, such as machine translation, named-entity recognition, and question answering, primarily due to its multi-head self-attention (MSA) mechanism, which establishes global connections between sequence tokens. The ability to model long-range dependencies also makes it suitable for pixel-based CV tasks. For instance, the Detection Transformer (DETR) [17] employs an elegant Transformer-based design to create the first fully end-to-end object detection model. Additionally, the Vision Transformer (ViT) [18], the first image recognition model purely based on the Transformer, has been proposed and achieved comparable performance with other SOTA convolution-based methods. A hierarchical Swin Transformer-V2 [19], [20] with Window-based MSA (W-MSA) and Shifted Window-based MSA (SW-MSA) has been proposed to reduce computational complexity, surpassing previous SOTA methods in image classification, dense prediction tasks such as object detection and semantic segmentation.

Based on the success of the Swin Transformer [20], we propose the STUCNet model to leverage the power of the Transformer for image segmentation of road cracks. STUCNet has a U-shaped architecture based on Transformer, CNN, and Attention including an encoder, a decoder, and skip connections. The encoder is built based on Swin Transformer-V2 blocks [20], and the decoder uses a CNN network like the basic UNET [11] adding an additional Attention mechanism. The input crack images are divided into non-overlapping image patches. Each patch is treated as a token and fed into the Transformer-based encoder to extract high-level feature representations. Contextual features are extracted and then upsampled by the decoder. It combines them with multi-scale features from the encoder through skip connections to restore the spatial resolution of the feature map and continue segmentation prediction. Experiments on road crack segmentation datasets demonstrate the proposed method's good segmentation accuracy.

Our contributions can be summarized as follows: Based on Swin Transformer-V2, CNN and Attention, we propose the STUCNet model with an Encoder - Decoder symmetric architecture with skip connections; In the encoder, local-to-global self-attention is performed; In the decoder, global features are upsampled according to the input resolution to predict corresponding pixel-level segmentation.

2. Background and related work

In this section, we summarize the most popular CNN-based methods used in crack image segmentation. Then, we provide an overview of recent research on Vision Transformers, focusing on their applications in segmentation tasks.

2.1. Road crack image segmentation based on CNNs

Detecting cracks at the pixel level for length and width determination is a semantic segmentation task. UNET [11] is a popular semantic segmentation network that utilizes an Encoder-Decoder architecture and skip-connections to concatenate shallow and deep feature maps, significantly enhancing the network performance. However, this approach increases the model parameters. SegNet [12] also employs skip-connections, allowing the maximum pooling parameter in the Encoder module to be introduced into the Decoder module, resulting in improved computational efficiency at the expense of detection accuracy. In 2017, CrackNet [13] was proposed for pixel-level crack detection tasks, which differs from traditional CNN structures by removing the pooling layer to avoid excessive downsampling that causes loss of detail. While CrackNet achieves better crack detection accuracy, artificially designed filters in CrackNet limit the learning ability of the model. An improved version of CrackNet, CrackNet-V [21], has a deeper structure, and fewer parameters. CrackNet-V has also improved crack detection and calculation efficiency. However, the algorithm's dataset only contains specific cracks, making it challenging to ensure the model's accuracy when the crack types change. Liu et al. [14] presented a model named DeepCrack, which detected cracks at the pixel level. The effectiveness of this model was confirmed through experiments on a dataset comprising 537 crack images. Although the above algorithms have achieved good results on their respective datasets, the datasets are quite small, making it difficult to prove the model's generalization ability. Li et al. [15] designed a crack segmentation model called SCCDNet, which helps to improve the effectiveness of segmentation and significantly reduces the model's parameters.

2.2. Vision Transformer

Recently, inspired by the great success of Transformers in natural language processing (NLP) [16], researchers have applied Transformer techniques to the field of computer vision [17]. In [18], the ViT was proposed to perform image recognition tasks. By taking 2D image patches with positional embeddings as input and pre-training on large datasets, ViT achieved performance comparable to CNN-based methods. Additionally, the Data-efficient image Transformer (DeiT) presented in [22] showed that Transformers can be trained on medium-sized datasets and can obtain a more powerful Transformer by combining it with feature extraction methods. In [19], [20], a hierarchical Swin Transformer-V2 was developed and improved. Using Swin Transformer-V2 as a feature extractor model, the authors of [20] achieved the highest performance in image classification, object detection, and semantic segmentation. The success of ViT, DeiT, and Swin Transformer in image recognition tasks demonstrates the potential applications of Transformers in the field of computer vision.

3. The proposed method

3.1. Overview of architecture

The overall architecture of the proposed STUCNet model is illustrated in Fig.1. STUCNet consists of an encoder, a decoder, and skip connections. The basic unit of the encoder in STUCNet is Swin Transformer-V2 blocks [20]. While in the decoder, it is a combination of CNN and Attention. For the encoder, to transform the inputs into sequence embeddings, crack images are divided into non-overlapping arrays with a size of 4×4. Using this partitioning approach, the feature dimension of each patch becomes $4 \times 4 \times 3 = 48$. Moreover, a linear embedding layer is applied to project the feature dimension into an arbitrary dimension (represented as C). The transformed patch tokens pass through several Swin Transformer-V2 blocks and patch merging layers to generate hierarchical feature representations. Specifically, the patch merging layer is responsible for downsampling and increasing dimension, while the Swin Transformer-V2 block is responsible for feature representation learning. The decoder includes an up-sampling operation, attention to identifying useful pixels, and CNN to learn feature representations. The extracted features are fused with multiscale features from the encoder via skip connections to complement the loss of spatial information caused by multiple downsampling. Finally, a combination of convolution and up-sampling is applied four times to restore the resolution of the object mapping to the input

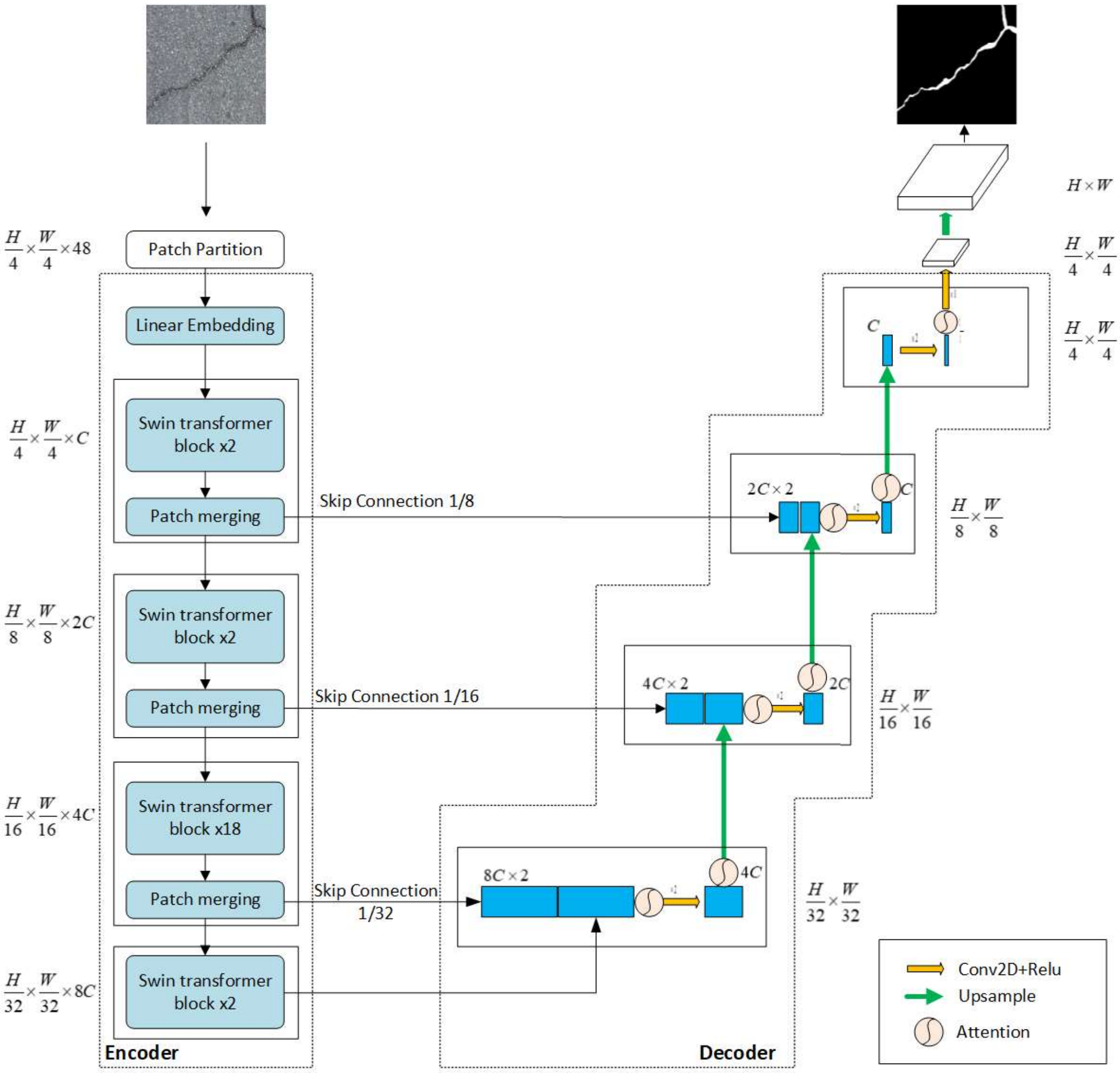
Patch Partition
Linear Embedding
Swin transformer block x2
Patch merging
Swin transformer block x2
Patch merging
Swin transformer block x18
Patch merging
Swin transformer block x2
Encoder
Decoder
Skip Connection 1/8
Skip Connection 1/16
Skip Connection 1/32
$\frac{H}{4}\times\frac{W}{4}\times 48$
$\frac{H}{4}\times\frac{W}{4}\times C$
$\frac{H}{8}\times\frac{W}{8}\times 2C$
$\frac{H}{16}\times\frac{W}{16}\times 4C$
$\frac{H}{32}\times\frac{W}{32}\times 8C$
$8C\times 2$
$4C$
$4C\times 2$
$2C$
$2C\times 2$
C
$H\times W$
$\frac{H}{4}\times\frac{W}{4}$
$\frac{H}{4}\times\frac{W}{4}$
$\frac{H}{8}\times\frac{W}{8}$
$\frac{H}{16}\times\frac{W}{16}$
$\frac{H}{32}\times\frac{W}{32}$
Conv2D+Relu
Upsample
Attention

Fig. 1. The overall architecture of STUCNet.

resolution (W × H), and then a linear projection layer is applied to output the pixel-level segmentation predictions.

3.2. Swin Transformer-V2 blocks

Unlike typical MSA input modules, the Swin Transformer-V2 [20] block is built based on shifted windows. Fig.2 illustrates two consecutive Swin Transformer-V2 blocks.

Each Swin Transformer-V2 block consists of a MSA module, a LayerNorm (LN) layer, residual connections, and 2 MLP layers with the GELU non-linear activation function. The first MSA module is based on a window (W-MSA). The second MSA

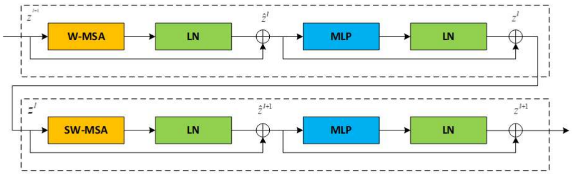

Fig. 2. Swin Transformer-V2 block.

module is based on shifted window (SW-MSA). These modules are applied in two consecutive transformer blocks. Based on such window partitioning mechanism, Swin Transformer-V2 blocks can be continuously constructed as:

$$\hat{z}^l = LN\left(W - MSA\left(z^{l-1}\right)\right) + z^{l-1}, \tag{1}$$

$$z^l = LN\left(MLP\left(\hat{z}^l\right)\right) + \hat{z}^l, \tag{2}$$

$$\hat{z}^{l+1} = LN\left(SW - MSA\left(z^l\right)\right) + z^l, \tag{3}$$

$$z^{l+1} = LN\left(MLP\left(\hat{z}^{l+1}\right)\right) + \hat{z}^{l+1}. \tag{4}$$

where $\hat{z}^l$ and z^l represent the outputs of the $(S)W-MSA$ module and the MLP module of the l^{th} block, respectively. Similar to the previous works [19], [20], self-attention is computed as follows:

$$Attention(Q, K, V) = SoftMax\left(\frac{QK^T}{\sqrt{d}} + B\right)V. \tag{5}$$

where $B \in \mathbb{R}^{M^2 \times M^2}$ is the relative position bias term for each head; $Q, K, V \in \mathbb{R}^{M^2 \times d}$ are the $query$, key and $value$ matrices; d is the $query/key$ dimension, and M^2 is the number of patches in a window.

We use 3 structures of Swin Transformer-V2 for the encoder of STUCNet with the following stage, block and channel settings:

- SwinV2-T: C = 96, block = $\{2; 2; 6; 2\}$
- SwinV2-S/B: C=96/128, block= $\{2; 2; 18; 2\}$

From there, the proposed STUCNet model also has three corresponding structures which are STUCNet-T, STUCNet-S, and STUCNet-B compared and evaluated for crack segmentation performance with other deep learning models in section 4.3.

In summary, each Swin transformer-V2 block has two mechanisms: the shift and the MSA. The shift mechanism is responsible for rearranging the position of pixels in the image, creating smaller patches that can be more efficiently processed by self-attention operations. The self-attention mechanism allows the model to weigh the importance of each pixel in the image against all other pixels. So, the model can capture long-range dependencies and contextual information. This helps the model better understand the relationships between different pixels in an image, which is particularly important for image segmentation tasks.

3.3. Encoder

In the encoder, the input is encoded into a $\frac{H}{4} \times \frac{W}{4}$ C-dimensional feature map. It is passed through two consecutive Swin Transformer-V2 blocks to perform feature extraction. In these blocks, the dimension and resolution of the features remain unchanged. At the same time, the Patch Merging layer reduces the number of tokens (2× downsampling) and doubles the dimension of the features. This process is repeated four times in the encoders. At the third iteration, it consists of 6 Swin Transformer-V2 blocks for the tiny architecture and 18 blocks for the small or base architectures.

Patch Merging Layer: The input patches are divided into four parts and merged by the patch merging layer. With such processing, the resolution of the features will be downsampled by a factor of 2. When merging patches leads to a 4x increase in feature size, a linear layer is applied to the merged features to unify the feature size to twice the original size.

Using the Swin transformer-V2 structure in the encoder module helps to improve model efficiency. This is because of Swin Transformer-V2's hierarchical architecture with repeated stages of transformer layers, reducing the number of parameters while maintaining high accuracy. It leads to faster and more efficient feature extraction compared to using convolutional blocks in the basic UNET network [11]. Additionally, Swin transformer-V2 also enhances model performance as it has worked well on large datasets with high-resolution input images.

3.4. Decoder

The decoder module interpolates features with multi-scale features from the encoder via skip connections. The module then feeds them into the first Attention layer. Hence, the combined model can selectively aggregate relevant features from the encoding block with those from the preceding decoding block to produce high-resolution output. Two feature selection convolutional operations are the same in UNET [11]. Finally, a second Attention layer is applied to attend to the informative pixels before passing to the next decoding block.

We apply 2 Attention layers in each decoding block to improve the segmentation accuracy. Because the attention mechanism in the decoding block helps the model learn to assign higher weights to more relevant features while suppressing or even ignoring irrelevant ones. In this way, the model can better capture the spatial relationships between different image features, leading to more detailed and accurate segmentation results.

3.5. Skip connection

Similar to UNET [11], skip connections are used to merge multi-scale features from the encoder with the upsampled features in the decoder. We combine shallow features and deep features together to minimize spatial information loss caused by downsampling. A linear layer follows, with the size of the concatenated features remaining the same as the size of the upsampled features. As the decoder has access to features from multiple scales, it can better identify and segment objects with different sizes and shapes. Skip connections also help reduce the number of parameters in the network, making training easier and avoiding model over-fitting.

4. Results and discussion

4.1. Dataset

Currently, there is no consistent publicly available dataset for crack segmentation tasks, making it difficult to compare different models on the same dataset. Most previous articles have trained and tested their models on small datasets they created themselves. Therefore, in this article, we created a large Crack-datasets by synthesizing crack images from various road surface crack datasets such as CFD, Deepcrack, Crack500, GAPS, Rissbilder, etc., to evaluate the effectiveness of different models. The dataset is publicly available at the following link: https://github.com/tungnlh/datasets. This dataset includes 7,169 manually labeled images with a resolution of 448 × 448. Some typical crack images and their labels are shown in Fig.3. It can be seen that the dataset contains crack images under various environmental conditions, different shooting distances, and different forms, including common and diverse crack characteristics.

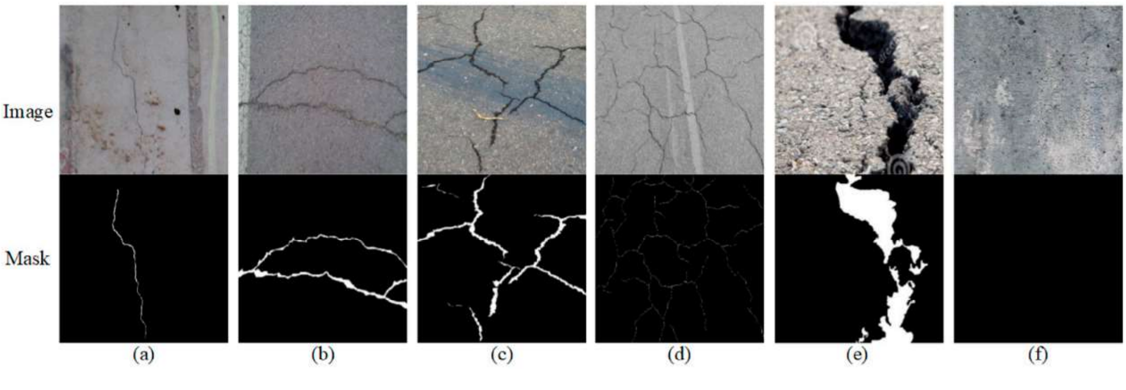

Fig. 3. Representative samples in the dataset. (a) A single crack in a complex environment, (b) A compound crack in a simple environment, (c) A crack taken at an oblique angle, (d) A net-like crack, (e) A coarse crack taken at close range, and (f) An image without crack.

The dataset is divided into two sets: the training set and the test set. The training set contains 6,164 images. In this set, 4,965 images have cracks, and 1,199 images do not have cracks. The test set involves 1,005 images which include 793 images with cracks and 212 images without cracks. The dataset has approximately 20% number of images without cracks to enhance the model's noise resistance and reduce the likelihood of false detection of environmental objects as cracks.

4.2. Evaluation matrix

Choosing appropriate evaluation metrics is necessary to quantitatively assess crack segmentation accuracy. To begin with, in crack detection, the following evaluation metrics are utilized:

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives}, \tag{6}$$

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives}, \tag{7}$$

$$F - score = \frac{2 \times Precision \times Recall}{Precision + Recall}. \tag{8}$$

The precision metric indicates the percentage of pixels that are accurately identified as cracks, revealing the influence of false positives on the outcome. On the other hand, the recall metric indicates the percentage of actual crack pixels that are correctly identified, showing the impact of false negatives on the results. The F-score metric is a combined measure of the two, taking into account the model's ability to minimize both false positives and false negatives. It provides a comprehensive evaluation of the model's ability to resist missed detections and false detections.

Furthermore, we present two widely used assessment metrics, Dice and Intersection over Union (IoU), in the field of semantic segmentation. The formula for these metrics is as follows:

$$Dice = \frac{2\left|\sum_i \sum_j X(i,j) \cdot Y(i,j)\right|}{\left|\sum_i \sum_j X(i,j)\right| + \left|\sum_i \sum_j Y(i,j)\right|}, \tag{9}$$

$$IoU = \frac{\left|\sum_i \sum_j X(i,j) \cdot Y(i,j)\right|}{\left|\sum_i \sum_j \left[(X(i,j) + Y(i,j)) - X(i,j) \cdot Y(i,j)\right]\right|}. \tag{10}$$

Not only is it essential to evaluate the segmentation accuracy of the model, but also its complexity, which serves as a crucial indicator. To measure the complexity, we utilize the model parameters and Floating Point Operations (FLOPs). The following formulas are employed to determine them:

$$Parameters = (D_k \times D_k \times M) \times N + N, \tag{11}$$

$$FLOPs = [(D_k \times D_k \times M) \times N + N] \times (H \times W) \,. \tag{12}$$

where D_k represents the size of the convolution kernel, M and N represent the number of channels of the input feature map and output feature map, respectively, and H and W represent the length and width of the output feature map, respectively.

4.3. Segmentation result

The STUCNet model can choose from 3 different structures of Swin transformer-V2 to be used as the encoder: tiny (T), small (S), and base (B). In addition, to further validate the advantages of STUCNet, we compared it with the segmentation methods DeepCrack [14], UNET [11], SegNet [12], and SCCDNet [15] on the same dataset. The comparison results are shown in Table.1.

Table 1. Segmentation results of different models. The best scores are in bold

Models	Precision	Recall	F-score	Dice	IoU
UNET	0.6953	0.8056	0.7464	0.7185	0.6002
SegNet	0.6483	0.7402	0.6912	0.6184	0.5015
DeepCrack	0.6761	0.4489	0.5396	0.3951	0.3166
SCCD-D32	0.7294	**0.8296**	0.7763	0.7541	0.6402
STUCNet-T	**0.8118**	0.7690	**0.7898**	**0.7717**	**0.6641**
STUCNet-S	0.7800	0.7844	0.7822	0.7602	0.6517
STUCNet-B	0.7940	0.7736	0.7836	0.7623	0.6538

From Table.1, it can be observed that the Tiny model achieved the highest accuracy in crack detection, with all metrics including Precision, F-score, Dice, and IoU achieving the highest results. The Small STUCNet-S and Base STUCNet-B models achieved similar results, and both outperformed modern crack segmentation models when trained and tested on the same relatively large and general dataset.

Compared to UNET, STUCNet-T achieved higher scores of 11.65%, 4.34%, 5.32%, and 6.39% on Precision, F-score, Dice, and IoU metrics, respectively. Even compared to SCCDNet, STUCNet-T outperformed with higher scores of 8.24%, 1.35%, 1.76%, and 2.39% on the same metrics, respectively. However, the Recall score was lower than both UNET and SCCDNet, indicating that the model's ability to avoid false negatives was not as good as SCCDNet. STUCNet-T performed better than STUCNet-S and STUCNet-B.

Regarding the complexity of the model, Table.2 compares the number of parameters and operations per second of each model. STUCNet-T has the lowest number of parameters among our three models and is only about 4.5M parameters higher than the best model SegNet. The STUCNet-B model has a relatively large number of parameters, about 2.8 times that of SegNet. The number of operations per second of STUCNet-T is also impressive, with only about 6 billion operations higher than SegNet and much lower than the other models.

Table 2. Compare the number of parameters of models, the number of operations per second, and the average time for segmenting an image

Models	FLOPs/G	Parameters/M	Time (ms/picture)
UNET	**111.632**	44.021	70
SegNet	**30.703**	**29.444**	59
DeepCrack	61.280	58.858	**33**
SCCD-D32	65.004	31.705	62
STUCNet-T	36.417	33.968	53
STUCNet-S	60.706	50.031	82
STUCNet-B	98.638	**81.526**	**109**

The average segmentation time per image is the mean value of each model's time to read, segment, and save 1,005 test images to jpg or png format. The results show that the STUCNet-T model (53 ms) is slower than DeepCrack (33 ms) but faster than the other models. This demonstrates that the segmentation speed of the model is quite good. The Small and Base structures of the model, due to their more complex structures, also have slower segmentation times, but the difference is not significant.

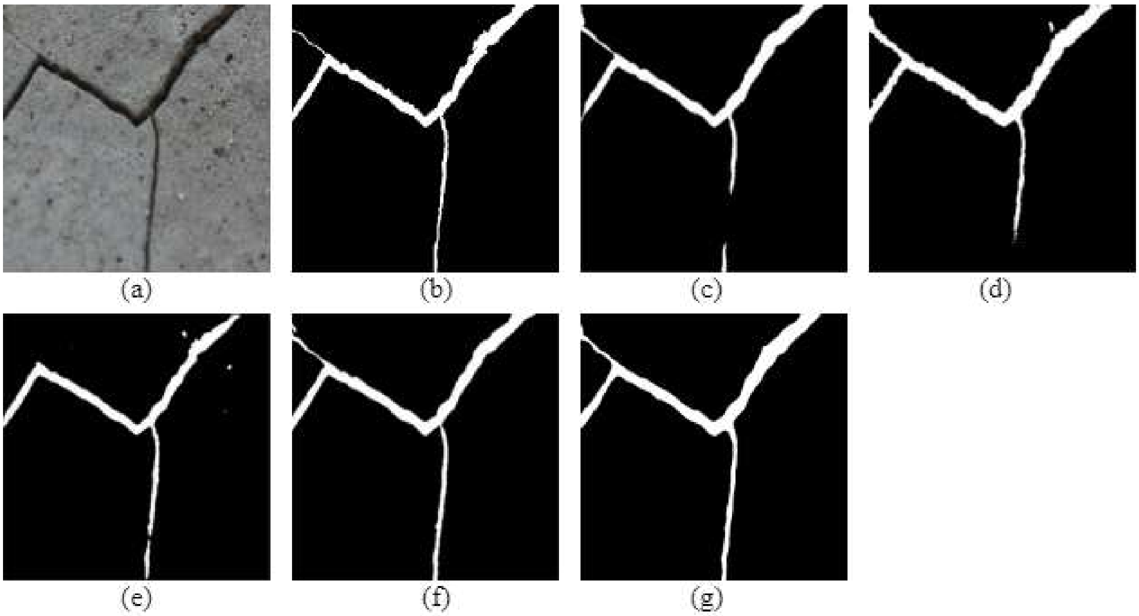

Fig. 4. Visualization of experimental results. (a) Input crack image, (b) Corresponding mask image, (c) Output from UNET, (d) Output from SegNet, (e) Output from DeepCrack, (f) Output from SCCD-D32, and (g) Output from STUCNet-T.

Fig.4 shows the experimental results of the typical models. It can be seen that STUCNet-T can detect all cracks on the input image, while other models only detect a part of the cracks or detect the cracks but with interruptions. SegNet and DeepCrack misidentify noise as cracks. This shows that our model has detected cracks more fully and improved its error detection ability. Although STUCNet gives good results, similar to SCCD-D32, it cannot detect some details of the structure on the edges of the crack.

This may be due to the window size in STUCNet still being large and sampling too much, leading to the loss of detailed features. We will research and design a more efficient structure to use detailed features on the feature map in the encoding module.

5. Conclusions

In this article, we introduce a STUCNet network model for crack segmentation that utilizes a symmetrical Encoder-Decoder architecture based on vision transformer, CNN, and Attention. The model consists of an encoding module with a Swin transformer-V2 network backbone with CNN, Attention, and skip connections to restore the spatial resolution of the feature map and continue segment prediction. The STUCNet model is evaluated on a large dataset of 7,169 labeled images collected from various environments. The results demonstrate that our model achieves higher accuracy than other advanced crack segmentation models in terms of precision, F-score, Dice, and IoU. In the future, we will focus on designing lighter and more advanced models to further improve crack segmentation results.

References

[1] H. Oliveira and P. L. Correia, “Automatic road crack segmentation using entropy and image dynamic thresholding,” in *2009 17th European Signal Processing Conference*. IEEE, 2009, pp. 622–626.

[2] R. Mishra, C. Chandrakar, and R. Mishra, “Surface defects detection for ceramic tiles using image processing and morphological techniques,” in *International*, vol. 2, no. 2, 2012.

[3] A. Zhang, Q. Li, K. C. Wang, and S. Qiu, “Matched filtering algorithm for pavement cracking detection,” *Transportation research record*, vol. 2367, no. 1, pp. 30–42, 2013.

[4] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998. doi: 10.1109/5.726791

[5] P. Sermanet, S. Chintala, and Y. LeCun, “Convolutional neural networks applied to house numbers digit classification,” *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pp. 3288–3291, 2012. doi: 10.48550/arXiv.1204.3968

[6] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90

[7] J. Redmon and A. Farhadi, “Yolo9000: Better, faster, stronger,” *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6517–6525, 2017. doi: 10.1109/CVPR.2017.690

[8] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431–3440, 2015. doi: 10.1109/CVPR.2015.7298965

[9] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, “Temporal segment networks: Towards good practices for deep action recognition,” in *European conference on computer vision*. Springer, 2016, pp. 20–36.

[10] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr, “Conditional random fields as recurrent neural networks,” *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1529–1537, 2015. doi: 10.1109/ICCV.2015.179

[11] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[12] V. Badrinarayanan, A. Kendall, and R. Cipolla, “Segnet: A deep convolutional encoder-decoder architecture for image segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017. doi: 10.1109/TPAMI.2016.2644615

[13] T. Yamane and P.-j. Chun, “Crack detection from a concrete surface image based on semantic segmentation using deep learning,” *Journal of Advanced Concrete Technology*, vol. 18, no. 9, pp. 493–504, 2020.

[14] Y. Liu, J. Yao, X. Lu, R. Xie, and L. Li, "Deepcrack: A deep hierarchical feature learning architecture for crack segmentation," *Neurocomputing*, vol. 338, pp. 139–153, 2019.

[15] H. Li, Z. Yue, J. Liu, Y. Wang, H. Cai, K. Cui, and X. Chen, "Sccdnet: A pixel-level crack segmentation network," *Applied Sciences*, vol. 11, no. 11, p. 5074, 2021.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[17] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer, 2020, pp. 213–229.

[18] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[19] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9992–10 002, 2021. doi: 10.1109/ICCV48922.2021.00986

[20] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, F. Wei, and B. Guo, "Swin transformer v2: Scaling up capacity and resolution," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11 999–12 009, 2022. doi: 10.1109/CVPR52688.2022.01170

[21] Y. Fei, K. C. P. Wang, A. Zhang, C. Chen, J. Q. Li, Y. Liu, G. Yang, and B. Li, "Pixel-level cracking detection on 3d asphalt pavement images through deep-learning- based cracknet-v," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 1, pp. 273–284, 2020. doi: 10.1109/TITS.2019.2891167

[22] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International conference on machine learning*. PMLR, 2021, pp. 10 347–10 357.

Manuscript received 15-03-2023; Accepted 12-06-2023.

■

Hai-Hong Phan received a Ph.D. degree in computer science from the University of CY Cergy Paris, France in 2019. She has more than 10 years of experience in research, teaching, consulting, and implementing information technology projects. She collaborates with high-ranking international journals as a reviewer. She has published scientific articles in high-quality journals and conferences such as IEEE Applied Sciences, Multimed Tools Appl, IET Image Processing, and International Conference on Pattern Recognition (ICPR). Her researches focus on computer vision, image processing, action recognition, and deep learning. E-mail: hongpth@lqdtu.edu.vn

Le Hoang Tung Nguyen graduated from the Army Engineering University of PLA (China) with a bachelor's degree in Computer science and Technology. Currently, he is doing his master's degree in Computer science at the Institute of information technology and Communication (IITC), Le Quy Don Technical University. His research interests are computer vision and AI. E-mail: tungnlh.bctt@gmail.com

MẠNG STUCNET NÂNG CAO HIỆU QUẢ NHẬN DẠNG VẾT ĐỨT MẶT ĐƯỜNG

Phan Hải Hồng, Nguyễn Lê Hoàng Tùng

Tóm tắt

Tự động phát hiện vết nứt mặt đường là một nhiệm vụ quan trọng nhằm hỗ trợ kiểm tra chất lượng đường bộ trong cơ sở hạ tầng giao thông. Nhiều phương pháp khác nhau đã được đề xuất để phân đoạn vết nứt mặt đường, tuy nhiên độ chính xác chưa cao. Để nâng cao hiệu quả nhận dạng vết nứt mặt đường, chúng tôi đã nghiên cứu đề xuất mô hình STUCNet (Swin Transformer-V2 UNET for Crack Segmentation Network) cho nhận dạng vết nứt mặt đường. Mô hình đề xuất kết hợp các ưu điểm của Swin Transformer-V2 vào mô-đun mã hóa của kiến trúc UNET tiêu chuẩn để nâng cao chất lượng phân đoạn ngữ nghĩa hình ảnh. Cụ thể, mô hình tích hợp Swin Transformer-V2 với các cửa sổ được dịch chuyển làm bộ mã hóa để trích xuất các đặc trưng ngữ cảnh cho phân đoạn vết nứt, bộ giải mã đối xứng dựa trên mạng nơ-ron tích chập cùng các Attention được thiết kế để thực hiện thao tác lấy mẫu lên nhằm khôi phục độ phân giải không gian và chú ý những đặc trưng quan trọng. Mô hình STUCNet được thực nghiệm trên bộ dữ liệu lớn chứa các vết nứt được thu thập trong các ngữ cảnh khác nhau. So với các mô hình tiên tiến hiện nay, mô hình của chúng tôi đạt được hiệu quả phân đoạn vết nứt tốt nhất.

Từ khóa

Nhận dạng vết nứt mặt đường, phân đoạn vết nứt, Transformer, Attention, UNET.