U-shaped Vision Transformer and Its Application in Gear Pitting Measurement

Sijun Wang, Yi Qin*, Dejun Xi, Chen Liang

College of Mechanical and Vehicle Engineering, Chongqing University, Chongqing, China

*Corresponding author: Yi Qin (e-mail: qy_808@cqu.edu.cn)

Abstract: Although convolutional neural networks (CNNs) have become the mainstream segmentation model, the locality of convolution makes them cannot well learn global and long-range semantic information. To further improve the performance of segmentation models, we propose u-shaped vision Transformer (UsViT), a model based on Transformer and convolution. Specifically, residual Transformer blocks are designed in the encoder of UsViT, which take advantages of residual network and Transformer backbone at the same time. What’s more, transpositions in each Transformer layer achieve the information interaction between spatial locations and feature channels, enhancing the capability of feature learning. In the decoder, for enhancing receptive filed, different dilation rates are introduced to each convolutional layer. In addition, residual connections are applied to make the information propagation smoother when training the model. We first verify the superiority of UsViT on Automatic Portrait Matting public dataset, which achieves 90.43% Acc, 95.56% DSC and 94.66% IoU with relatively fewer parameters. Finally, UsViT is applied to gear pitting measurement in gear contact fatigue test, and the comparative results indicate that UsViT can improve the accuracy of pitting detection.

Keywords: Vision transformer; residual connection; dilation rate; information interaction; pitting measurement

I. INTRODUCTION

Gear is a widely-used motion and power transmission component in mechanical equipment, and it is prone to failure due to the poor working condition. Moreover, pitting is the main failure mode of gear, which has been detected by vibration-based methods in the past years [1-3]. However, it is difficult for vibration-based methods to quantitatively detect gear pitting. Gear pitting area ratio is a key metric for evaluating the degree of failure, especially in the gear contact fatigue test [4]. In order to calculate gear pitting area ratio, machine vision may be a feasible tool [5]. Based on machine vision methods, the key to measuring gear pitting area ratio is the precise segmentation of the effective tooth surface and the pitting from the acquired gear image. However, the gear pitting is generally irregular, which bring great challenges to the traditional computer vision techniques including threshold segmentation and edge segmentation [6-8].

Convolutional neural networks (CNNs) have shown excellent feature extraction ability and good semantic segmentation performance in recent years, so they were successfully applied to different fields of automatic detection [9-11]. Zhang et al. [12] proposed a simple but efficient segmentation model for road area extraction by applying residual connection into U-Net. Ding et al. [13] proposed an improved algorithm based on the encoder-decoder
framework of U-Net to accurately segment the common defects such as untitled corner, scratch and dirty in the process of magnetic disc quality detection. Du et al [14] proposed a seismic crack recognition method based on ResU-Net and dense CRF model, improving the efficiency and accuracy in the detection of seismic image dataset. Li et al [15] proposed a ship detection method based on U-Net++ and multiple side-output fusion algorithm, solving the problems of complicated background and various ship sizes in satellite remote sensing images. However, there are few studies on the gear pitting detection, and the accurate segmentation of background, effective tooth surface and pitting area still face a big challenge. Also, it is worth noting that all the above methods are based on CNNs which cannot well learn global and long-range semantic information. To further improve the segmentation performance of gear pitting image, it is indispensable to develop a model which takes advantages of local and global feature information simultaneously.

Transformer [16] has made a great achievement in natural language processing (NLP) over the last few years because of its powerful ability of learning long-range feature information. ViT [17] was the first work to introduce Transformer into computer vision, which had an outstanding performance on image classification. Except for classification, Transformer was also applied to dense prediction tasks such as semantic segmentation and object detection. Zheng et al. [18] proposed segmentation Transformer (SETR), which provided two kinds of decoder and achieved better performance of semantic segmentation. PVT [19] realized multiscale outputs and low computational complexity by introducing a pyramid structure. P2T [20] enhanced the contextual information by adding the pyramid pooling to Transformer. SegFormer [21] proposed a pure MLP decoder to aggregate information from different layers. The excellent segmentation performance of the foregoing methods provides an idea to explore Transformer-based model for gear pitting measurement.

In this paper, we propose u-shaped vision Transformer (UsViT), an efficient and powerful framework for gear pitting measurement. Concretely, our contributions can be summarized as: (1) Residual Transformer blocks are designed in the encoder of UsViT, which take advantages of residual network and Transformer backbone at the same time. What’s more, transpositions in each Transformer layer achieve the information interaction between spatial locations and feature channels, so as to enhance the capability of feature learning (2) Different dilation rates are introduced to each convolutional layer for improving the receptive field in the decoder. In addition, the residual connections are applied to make the information propagation smoother when training the model (3) UsViT achieves better performance on Automatic Portrait Matting public dataset with relatively fewer parameters. Finally, UsViT is applied to gear pitting measurement in the gear contact fatigue test, and the comparative results indicate that UsViT can improve the accuracy of pitting detection.

II. METHOD

A. ARCHITECTURE OVERVIEW

As depicted in Fig. 1, there are two basic modules in UsViT: Transformer layer in the encoder and Decoder layer in the decoder. Given an input image with $H \times W \times 3$, UsViT first downsamples it into several non-overlapping patches of $\frac{H}{16} \times \frac{W}{16}$ by linear projection. Then, these patches are flattened, and input to the Transformer block as a sequence. Through several residual Transformer blocks, we can reshape the output feature map into $\frac{H}{16} \times \frac{W}{16} \times C$. Inspired by U-Net, we design a similarly progressive upsampling decoder including four same stages with $2 \times$ to reach the full resolution of $H \times W$. We feed the reshaped feature map into the decoder and there is a Decoder layer in each stage. In the end, the output feature map from the decoder is processed by a $1 \times 1$ convolutional layer with softmax activation.
function for predicting the pixel-level segmentation mask.

Fig. 1. The overall architecture of UsViT.

B. TRANSFORMER LAYER

Different from the original structure of Transformer encoder in ViT, the Transformer block in UsViT is constructed with the idea of residual connection, which consists of two Transformer layers. The input sequence is first normalized by Layer Norm (LN), then feed into multi-head self-attention (MSA). After added by the input sequence, the feature map is processed by transposition subsequently. Through the similar step of LN, MLP and transposition, we can obtain the output feature map of one Transformer layer. Two transpositions in Transformer layer achieve the information interaction between spatial locations and feature channels, enhancing the capability of feature information learning. The Transformer block can be formulated as:

\[ F_i = (MLP(LN(\hat{F}_{i-1}^t) + \hat{F}_{i-1}^t))^t \]  
\[ \hat{F}_{i} = (MSA(LN(F_{i-1})) + F_{i-1})^t \]  
\[ \hat{F}_{i+1}^t = (MSA(LN(F_i)) + F_i)^t \]  

where \( F_{i-1} \), \( F_i \) and \( F_{i+1} \) are the input or output feature map in a Transformer layer. \( \hat{F}_{i} \), \( \hat{F}_{i+1}^t \) are the output feature map of MSA module. \( F_o \) is the output feature map of a Transformer block.

In the Transformer layer, the multi-head self-attention [22-23] is formulated as:

\[ \text{Attention}(Q, K, V) = \text{Softmax} \left( \frac{QK^T}{\sqrt{d_k}} \right) W \]  

where \( Q, K, V \) are calculated by the product of three learnable parameters with the input feature maps; \( d_k \) denotes the channel dimension of \( K \).

C. DECODER LAYER

After several Transformer blocks, we can obtain fine-grained feature maps, which are subsequently reshaped into \( \frac{n}{16} \times \frac{w}{16} \times C \). The reshaped feature maps are then feed into a progressive upsampling decoder including four same stages with 2× to reach the full resolution of \( H \times W \). The details of resolution
and dimension of different stages are shown in Fig. 1. Each stage consists of a Decoder layer, which includes a transposed convolutional layer and 4 convolutional layers with residual connections. Furthermore, instead of using the usual convolutions, different dilation rates (1, 3, 5, 7) are introduced to each convolutional layer for enhancing receptive field. In addition, with the help of residual connections, multiscale feature information from different dilated convolutions are fused, thus the performance of semantic segmentation will be boosted. Finally, the output feature map from the decoder is processed by a 1 × 1 convolutional layer with softmax activation function for predicting the pixel-level segmentation mask.

III. EXPERIMENTS

A. DATASET

To verify the superior performance of UsViT, we first conduct experiments on a public dataset. Automatic Portrait Matting [24] is a portrait segmentation dataset collected from Flickr, which contains 2000 images with high-quality portraits. With the resolution of 800 × 600, these images are randomly split into 1500, 200 and 300 for training, validating and testing respectively. In addition, the labeling process is finished by closed-form [25] and KNN [26] matting to make sure the high quality of the dataset.

B. IMPLEMENTATION DETAILS

UsViT was trained by a computing platform with a NVIDIA GTX 2080Ti based on Python 3.6 and Tensorflow 2.1. Due to the limitation of computing resources, the input images were resized to 256 × 256. During the training period, Adam was used as the optimizer and the total epochs was set to 120. The initial value of learning rate was set to 0.0001 and the batch size is 8. In the experiment, we used CrossEntropy as the loss function. Accuracy (Acc), Dice-Similarity coefficient (DSC [27]) and Intersection over Union (IoU [28]) are used as evaluation metrics for the test set.

C. EXPERIMENTAL RESULTS

The comparison of the proposed UsViT model with U-Net and its variants on the Automatic Portrait Matting are shown in Table I. It is apparent from the table that UsViT achieves the best segmentation performance with 90.43% Acc, 95.56% DSC and 94.66% IoU. Compared to U-Net, UsViT gets 1.27% Acc improvement, 2.41% DSC improvement and 2.88% IoU improvement respectively. What’s more, UsViT yields the best performance with only 25.6M parameters while U-Net has 31.1M parameters. To sum up, UsViT has higher segmentation performance with relatively fewer parameters than other compared models.

Fig. 2 demonstrates some segmentation results of portrait images obtained by various segmentation models. As depicted in Fig. 2, the segmentation result of UsViT are closer to the ground truth compared to other models. Especially at the edge of the segmentation result, UsViT shows smooth and similar details as the ground truth. It then indicates that UsViT has stronger learning ability of feature representation and better segmentation performance. In a word, UsViT takes advantages of convolution and Transformer meanwhile, which can realize the interaction of local and global semantic information, so it can obtain better segmentation results.

Table I. Segmentation performance of Automatic Portrait Matting dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>Param</th>
<th>Acc</th>
<th>DSC</th>
<th>IoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>UsViT</td>
<td>25.6M</td>
<td>90.43</td>
<td>95.56</td>
<td>94.66</td>
</tr>
<tr>
<td>U-Net</td>
<td>31.1M</td>
<td>89.16</td>
<td>93.15</td>
<td>91.78</td>
</tr>
<tr>
<td>ResU-Net</td>
<td>67.4M</td>
<td>89.35</td>
<td>93.42</td>
<td>92.12</td>
</tr>
<tr>
<td>R2U-Net</td>
<td>70.5M</td>
<td>82.17</td>
<td>85.89</td>
<td>84.45</td>
</tr>
</tbody>
</table>
Fig. 2. Segmentation results of Automatic Portrait Matting dataset.

D. ABLATION STUDY

In this experiment, we conducted ablation study on Automatic Portrait Matting with reducing one factor at a time to investigate the contributions of different factors. There are four situations about base model: A1-without residual connections in the encoder; A2-without transpositions in the Transformer layer; A3-without residual connections in the Decoder layer; A4-without dilation rates in the Decoder layer.

Table II shows the results of ablation study. It is obvious that residual connections in the encoder and dilation rates in the Decoder layer make more contribution to the improvement of segmentation performance. Residual connections in the encoder make the feature information propagation smoother. Different dilation rates in the Decoder layer enhance receptive field and aggregate characteristic information. UsViT combines the advantages of both factors, thereby achieving the best segmentation performance.

Table II. Ablation study on the impact of different factors

<table>
<thead>
<tr>
<th>Ablation of UsViT</th>
<th>Acc</th>
<th>DSC</th>
<th>IoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base model</td>
<td>90.43</td>
<td>95.56</td>
<td>94.66</td>
</tr>
<tr>
<td>A1</td>
<td>89.14</td>
<td>93.23</td>
<td>92.37</td>
</tr>
<tr>
<td>A2</td>
<td>90.25</td>
<td>94.27</td>
<td>93.52</td>
</tr>
<tr>
<td>A3</td>
<td>90.17</td>
<td>94.13</td>
<td>93.28</td>
</tr>
<tr>
<td>A4</td>
<td>88.89</td>
<td>93.61</td>
<td>92.46</td>
</tr>
</tbody>
</table>

IV. APPLICATION

A. ACQUISITION OF GEAR PITTING IMAGES

As shown in Fig. 3, the first step in gear pitting measurement is to acquire gear pitting images. The experimental device is presented in Fig. 4. The left of Fig. 4 shows general view of test rig, while the right illustrates the test gearbox and the vision measuring device. For online acquiring the image of gear teeth, a vision measuring system was designed. First, we used a transparent plexiglass plate as upper cover of test gearbox for clearly taking photographs of gear teeth. Then, to facilitate the adjustment of the shooting angle, the CCD industrial camera was fixed in a flexibly adjustable bracket, and LED light source was used. Via gear contact fatigue experiments and the vision measuring system, 800 gear pitting images were collected. To make complete gear pitting dataset, we made corresponding labels by Labelme image annotation tool. The resolution of pitting images is 256×256, and the ratio of the training images’ quantity, the validating images’ quantity and the testing images’ quantity is 7:1:2. Except that the iteration is set to 160 on Gear Pitting dataset, other implementation details is the same as that of Automatic Portrait Matting dataset.

![Flow chart of application in gear pitting measurement.](image)

B. RELATIVE ERROR

In addition to the three metrics of Acc, DSC and IoU, the relative error of gear pitting area ratio (Re) is also employed to evaluate segmentation
performance on the gear pitting dataset. After counting the number of pixels from effective tooth surface area \( (A_e) \) and pitting area \( (A_p) \) in acquired image, the gear pitting area ratio \( (AR) \) can be calculated by

\[
AR = \frac{A_p}{A_e} \times 100\%
\]

(7)

After the actual pitting area ratio \( (AR_a) \) and the predicted pitting area ratio \( (AR_p) \) are calculated, \( Re \) can be computed by the following formula. Additionally, average \( Re \) (i.e. \( Re \)) of multiple test images is used to compare segmentation performance of different models.

\[
Re = \left| \frac{AR_a - AR_p}{AR_a} \right| \times 100\%
\]

(8)

Test results of Gear Pitting dataset are listed in Tab respectively, which indicates that our approach can achieve better segmentation performance and save calculation cost. What’s more, \( Re \) (average \( Re \)) of UsViT is only 6.78%, which is smaller than those of other segmentation models. Therefore, the proposed UsViT is more suitable for calculating the gear pitting area ratio.

Fig. 5 demonstrates some segmentation results of \( g \) takes advantages of convolution and Transformer meanwhile, which can realize the interaction of local and global semantic information, resulting in better segmentation results.

| Table III. Segmentation performance of Gear Pitting dataset |
|-----------------|-------|------|------|-------|
| Method          | Param | Acc  | DSC  | mIoU  | Average Re |
| UsViT           | 25.6M | 97.91| 94.52| 92.45 | 6.78      |
| U-Net           | 31.1M | 96.73| 93.19| 90.98 | 8.94      |
| ResU-Net        | 67.4M | 96.24| 92.45| 90.62 | 9.83      |
| R2U-Net         | 70.5M | 95.86| 91.79| 84.21 | 14.65     |
**CONCLUSION**

In this paper, we propose UsViT, an efficient and powerful framework based on Transformer and convolution. Specifically, residual Transformer blocks are designed in the encoder of UsViT, which take advantages of residual network and Transformer backbone meanwhile. What’s more, transpositions in each Transformer layer achieve the information interaction between spatial locations and feature channels, enhancing the capability of feature learning. In the decoder, for enhancing receptive filed, different dilation rates are introduced to each convolutional layer. In addition, residual connections are applied to make the information propagation smoother when training the model. The experiments on Automatic Portrait Matting public dataset verify the advantages of the proposed UsViT, which achieves 90.43% Acc., 95.56% DSC and 94.66% IoU with relatively fewer parameters. Finally, UsViT is applied to gear pitting measurement in gear contact fatigue test, and the comparative results indicate that UsViT can improve the accuracy of pitting detection.

**ACKNOWLEDGMENTS**

This work was supported in part by National Natural Science Foundation of China under Grants 62033001 and 52175075.

**REFERENCES**


