Journal of Dynamics, Monitoring and Diagnostics, 2022, 1, 223-228
https://doi.org/10.37965/jdmd.2022.130

ISTP
RESEARCH ARTICLE

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

Sijun Wang,[1] Yi Qin,[1] Dejun Xi,[1] and Chen Liang[1]
[1]College of Mechanical and Vehicle Engineering, Chongqing University, Chongqing, China

(Received 23 August 2022; Revised 14 October 2022; Accepted 24 November 2022; Published online 24 November 2022)

Abstract: Although convolutional neural networks 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 is 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 field, 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% accuracy (Acc), 95.56% Dice similarity coefficient, and 94.66% Intersection over Union 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 Acc 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 untilled 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 (Acc) 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 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, which provided two kinds of decoder and achieved better performance of semantic segmentation. Pyramid Vision Transformer (PVT) [19] realized multiscale outputs and low computational complexity by introducing a pyramid structure. Pyramid Pooling Transformer (P2T) [20] enhanced the contextual information by adding the pyramid pooling to Transformer. SegFormer [21] proposed a pure Multilayer Perceptron (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

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

designed in the encoder of UsViT, which take advantages of residual network and Transformer backbone at the same time. What is 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 Acc 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× 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 1×1 convolutional layer with softmax activation function for predicting the pixel-level segmentation mask.

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) and then feeds 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:

$$\widehat{F}_i = (MSA(LN(F_{i-1})) + F_{i-1})^T \tag{1}$$

$$F_i = (MLP(LN(\widehat{F}_i) + \widehat{F}_i))^T \tag{2}$$

$$\widehat{F_{i+1}} = (MSA(LN(F_i)) + F_i)^T \tag{3}$$

$$F_{i+1} = (MLP(LN(\widehat{F_{i+1}}) + \widehat{F_{i+1}}))^T \tag{4}$$

$$F_o = F_{i+1} + F_{i-1} \tag{5}$$

where F_{i-1}, F_i, and F_{i+1} are the input or output feature map in a Transformer layer. $\widehat{F}_i$, $\widehat{F_{i+1}}$ are the output feature map of MSA module. F_o is the output feature map of a Transformer block.

In the Transformer layer, the MSA [22,23] is formulated as:

$$\text{Attention}(Q,K,V) = Softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{6}$$

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.

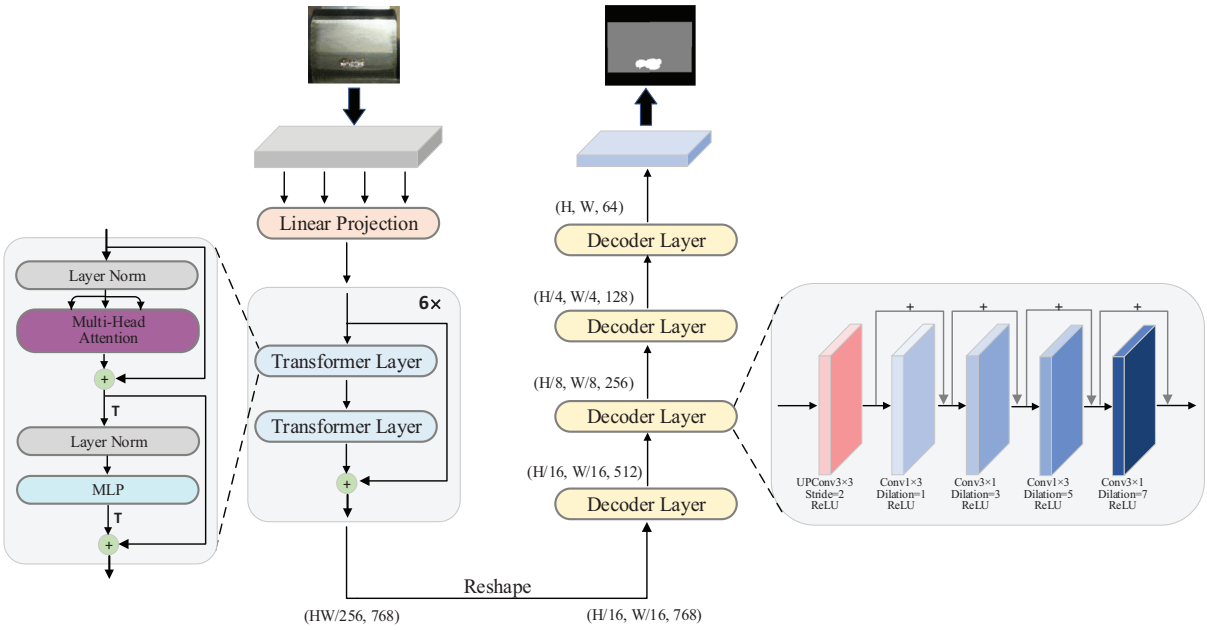

Fig. 1. The overall architecture of UsViT.

C. DECODER LAYER

After several Transformer blocks, we can obtain fine-grained feature maps, which are subsequently reshaped into $\frac{H}{16} \times \frac{W}{16} \times C$. The reshaped feature maps are then fed 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 is 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 K-Nearest-Neighbor (KNN) [26] matting to make sure the high quality of the dataset.

B. IMPLEMENTATION DETAILS

UsViT was trained by a computing platform with a NIVIDIA 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 were 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 cross entropy as the loss function. 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 is more, UsViT yields the best performance with only 25.6 M parameters while U-Net has 31.1 M parameters. To sum up, UsViT has higher segmentation performance with relatively fewer parameters than other compared models.

Table I. Segmentation performance of automatic portrait matting dataset

Method	Param	Acc	DSC	IoU
UsViT	**25.6 M**	**90.43**	**95.56**	**94.66**
U-Net	31.1 M	89.16	93.15	91.78
ResU-Net	67.4 M	89.35	93.42	92.12
R2U-Net	70.5 M	82.17	85.89	84.45

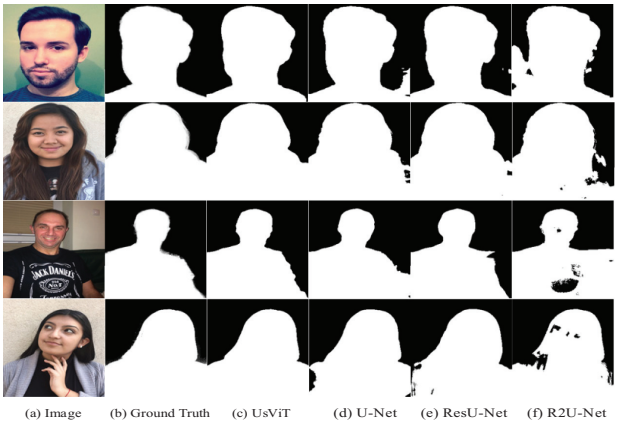

Fig. 2. Segmentation results of automatic portrait matting dataset.

Figure 2 demonstrates some segmentation results of portrait images obtained by various segmentation models. As depicted in Fig. 2, the segmentation result of UsViT is 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.

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; and 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

Table II. Ablation study on the impact of different factors

Ablation of UsViT	Acc	DSC	IoU
Base model	90.43	95.56	94.66
A1	89.14	93.23	92.37
A2	90.25	94.27	93.52
A3	90.17	94.13	93.28
A4	88.89	93.61	92.46

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.

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 Charge Coupled Device (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 are the same as that of automatic portrait matting dataset.

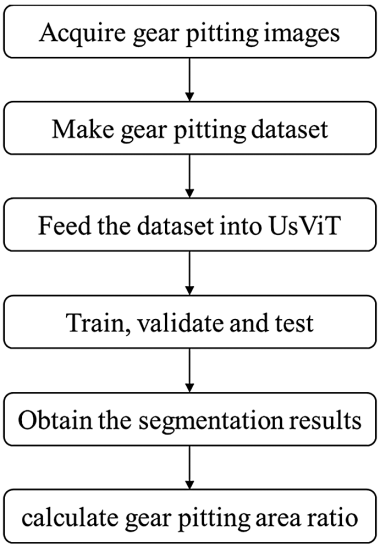

Fig. 3. Flow chart of application in gear pitting measurement.

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_t) and pitting area (A_p) in acquired image, the gear pitting area ratio (AR) can be calculated by

$$AR = \frac{A_p}{A_t} \times 100\% \tag{7}$$

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. $\overline{\text{Re}}$) of multiple test images is used to compare segmentation performance of different models.

$$\text{Re} = \left| \frac{AR_p - AR_a}{AR_a} \right| \times 100\% \tag{8}$$

C. EXPERIMENT RESULTS

Test results of gear pitting dataset are listed in Table III. From the table, we can note that the segmentation performance of

Table III. Segmentation performance of gear pitting dataset

Method	Param	Acc	DSC	mIoU	Average Re
UsViT	**25.6 M**	**97.91**	**94.52**	**92.45**	**6.78**
U-Net	31.1 M	96.73	93.19	90.98	8.94
ResU-Net	67.4 M	96.24	92.45	90.62	9.83
R2U-Net	70.5 M	95.86	91.79	84.21	14.65

(a) (b)

Fig. 4. Test rig and vision measuring system for gear contact fatigue. (a) test rig (b) vision measuring system.

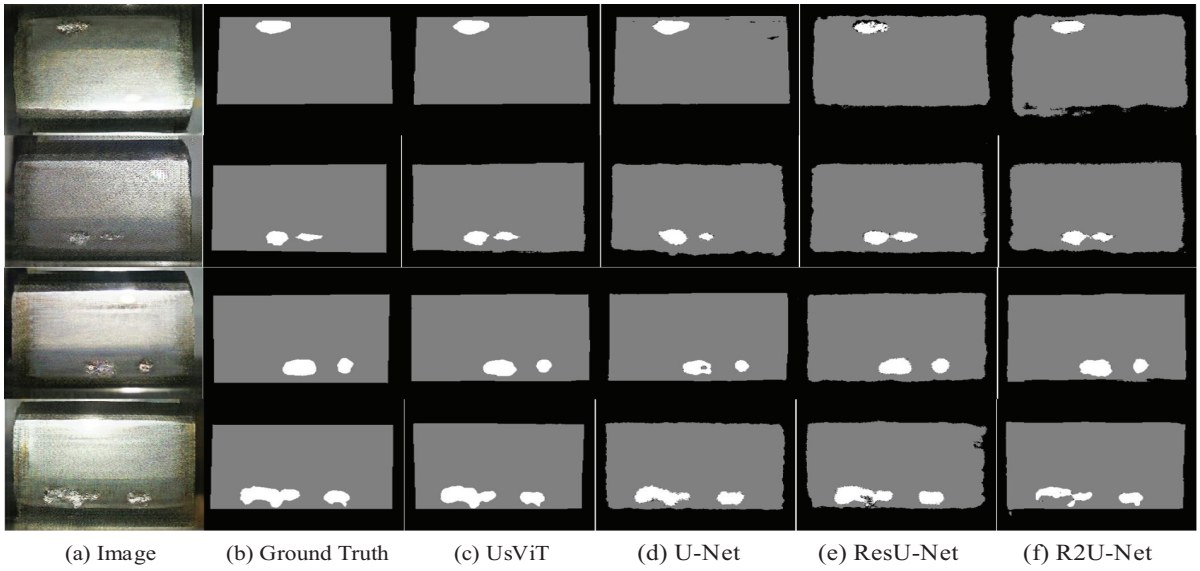

Fig. 5. Segmentation results of gear pitting dataset.

UsViT is better than other methods with 97.91% Acc, 94.52% DSC, and 92.45% mIoU (average IoU) at relatively fewer training parameters. Compared to U-Net, UsViT gets 1.18% Acc improvement, 1.33% DSC improvement, and 1.47% IoU improvement, respectively, which indicates that our approach can achieve better segmentation performance and save calculation cost. What is more, $\overline{\text{Re}}$ (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.

Figure 5 demonstrates some segmentation results of gear pitting images obtained by various segmentation models. As depicted in Fig. 5, the pitting and effective tooth surface segmented by UsViT are better, compared to other models. It then indicates that UsViT has stronger learning ability of feature representation and better segmentation performance for small targets. Therefore, UsViT takes advantages of convolution and Transformer meanwhile, which can realize the interaction of local and global semantic information, resulting in better segmentation results.

V. 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 is 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 field, 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 Acc of pitting detection.

Acknowledgments

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

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

References

[1] Y. Xu, X. Tang, G. Feng, D. Wang, C. Ashworth, F. Gu, and A. Ball, "Orthogonal on-rotor sensing vibrations for condition monitoring of rotating machines," *J. Dyn. Monit. Diagn.*, vol. 1, pp. 29–36, 2022.

[2] W. Wang, Y. Lei, T. Yan, N. Li, and A. Nandi, "Residual convolution long short-term memory network for machines remaining useful life prediction and uncertainty quantification," *J. Dyn. Monit. Diagn.*, vol. 1, pp. 2–8, 2022.

[3] B. Chen, D. Song, Y. Cheng, W. Zhang, B. Huang, and Y. Muhamedsalih, "IGIgram: an improved Gini index-based envelope analysis for rolling bearing fault diagnosis," *J. Dyn. Monit. Diagn.*, vol. 1, pp. 111–124, 2022.

[4] D. Xi, Y. Qin, J. Luo, H. Pu, and Z. Wang, "Multipath fusion mask R-CNN with double attention and its application into gear pit-ting detection," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–11, 2021.

[5] Y. Qin, Z. Wang, and D. Xi, "Tree CycleGAN with maximum diversity loss for image augmentation and its application into gear pitting detection," *Appl. Soft Comput.*, vol. 114, pp. 108–130, 2022.

[6] W. Zhang, X. Wang, W. You, J. Chen, P. Dai, and P. Zhang, "RESLS: region and edge synergetic level set framework for image segmentation," *IEEE Trans. Image Process.*, vol. 29, pp. 57–71, 2020.

[7] X. Zhang, D. Rajan, and B. Story., "Concrete crack detection using context-aware deep semantic segmentation network," *Comput.-Aided Civ. Infrastruct. Eng.*, vol. 34, pp. 951–971, 2019.

[8] S. Meister, M. Wermes, J. Stüve, and R. Groves, "Review of image segmentation techniques for layup defect detection in the Automated Fiber Placement process: a comprehensive study to improve AFP inspection," *J. Intell. Manuf.*, vol. 32, pp. 2099–2119, 2021.

[9] Y. Wang, Y. Ni, and X. Wang, "Real-time defect detection of high-speed train wheels by using Bayesian forecasting and dynamic model," *Mech. Syst. Signal Process.*, vol. 139, 2020.

[10] Y. He, K. Song, Q. Meng, and Y. Yan, "An End-to-End Steel Surface Defect Detection Approach via Fusing Multiple Hierarchical Features," *IEEE Trans. Instrum. Meas.*, vol. 69, pp. 1493–1504, 2020.

[11] S. Zhang, Q. Zhang, J. Gu, L. Su, K. Li, and M. Pecht, "Visual inspection of steel surface defects based on domain adaptation and adaptive convolutional neural network," *Mechanical Systems and Signal Processing*, vol. 153, pp. 447–457, 2021.

[12] Z. Zhang, Q Liu, and Y Wang, "Road extraction by deep residual U-net," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 749–753, 2018.

[13] Y. Ding and T. Hong, "Detection of surface defects of magnetic sheets based on improved U-net convolutional neural network," *China Standardization*, vol. 13, pp. 192–198, 2021.

[14] C. Du, Y. Duan, and Q. Sun, "A seismic crack recognition method based on ResUNet and dense CRF model," *J. Appl. Sci.*, vol. 39, no. 3, pp. 367–377, 2021.

[15] Z. Li, H. Yin, J. Zuo, and Y. Sun, "Ship detection method based on UNet++ and multilateral output fusion algorithm," *Comput. Eng.*, vol. 48, pp. 276–283, 2021.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, and L. Kaiser, "Attention is all you need", *ArXiv Preprint ArXiv: 1706.03762*, 2017.

[17] A. Dosovitskiy et al., "An image is worth 16x16 words: transformers for image recognition at scale," *ArXiv Preprint ArXiv: 2010.11929v2*, 2020.

[18] S. Zheng et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recogn.*, pp. 6881–6890, 2021.

[19] W. Wang et al., "Pyramid vision transformer: a versatile backbone for dense prediction without convolutions," *Proc. IEEE/CVF Conf. Comput. Vision*, pp. 548–558, 2021.

[20] Y. Wu, Y. Liu, X. Zhan, and M. Cheng, "P2T: pyramid pooling transformer for scene understanding," *ArXiv Preprint ArXiv: 2106.12011v1*, 2021.

[21] E. Xie, W. Wang, Z. Yu, A. Anandkumer, J. Alvarez, and P. Luo, "SegFormer: simple and efficient design for semantic segmentation with transformers," *ArXiv Preprint ArXiv: 2105.15203v1*, 2021.

[22] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei, "Relation networks for object detection," *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recogn.*, pp. 3588–3597, 2018.

[23] H. Hu, Z. Zhang, Z. Xie, and S. Lin, "Local relation networks for image recognition," *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, pp. 3463–3472, 2019.

[24] X. Shen, X. Tao, H. Gao, C. Zhou, and J. Jia, "Deep automatic portrait matting," *Eur. Conf. Comput. Vision*, pp. 92–107, 2016.

[25] A. Levin, D. Lischinski, Y. Weiss, "A closed-form solution to natural image matting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp 228–242, 2008.

[26] Q. Chen, D. Li, and C. Tang, "KNN matting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 9, pp. 2175–2188, 2013.

[27] L. Luo, D. Xue, and X. Feng, "Automatic segmentation of retinal blood vessels based on compact hybrid network," *Control Decis.*, vol. 37, pp. 353–360, 2021.

[28] N. Ibtehaz, M. Rahman,MultiResUNet, "Rethinking the U-net architecture for multimodal biomedical image segmentation," *Neural Netw.*, vol. 121, pp. 74–87, 2020.