Qin *et al.*
Chinese Journal of Mechanical Engineering (2023) 36:50
https://doi.org/10.1186/s10033-023-00874-w

Chinese Journal of Mechanical Engineering

ORIGINAL ARTICLE

Open Access

Gear Pitting Measurement by Multi-Scale Splicing Attention U-Net

Yi Qin[1,2*], Dejun Xi[1,2], Weiwei Chen[1,2] and Yi Wang[1,2]

Abstract

The judgment of gear failure is based on the pitting area ratio of gear. Traditional gear pitting calculation method mainly rely on manual visual inspection. This method is greatly affected by human factors, and is greatly affected by the working experience, training degree and fatigue degree of the detection personnel, so the detection results may be biased. The non-contact computer vision measurement can carry out non-destructive testing and monitoring under the working condition of the machine, and has high detection accuracy. To improve the measurement accuracy of gear pitting, a novel multi-scale splicing attention U-Net (MSSA U-Net) is explored in this study. An image splicing module is first proposed for concatenating the output feature maps of multiple convolutional layers into a splicing feature map with more semantic information. Then, an attention module is applied to select the key features of the splicing feature map. Given that MSSA U-Net adequately uses multi-scale semantic features, it has better segmentation performance on irregular small objects than U-Net and attention U-Net. On the basis of the designed visual detection platform and MSSA U-Net, a methodology for measuring the area ratio of gear pitting is proposed. With three datasets, experimental results show that MSSA U-Net is superior to existing typical image segmentation methods and can accurately segment different levels of pitting due to its strong segmentation ability. Therefore, the proposed methodology can be effectively applied in measuring the pitting area ratio and determining the level of gear pitting.

Keywords Gear pitting, Image segmentation, Attention module, Computer vision, Quantitative detection

1 Introduction

Gear is an extremely important core basic component of high-end equipment, and it is also one of the most failure-prone transmission components. Therefore, in order to ensure the normal and safe operation of major equipment, it is necessary to carry out the fault detection research of gear system. Gear pitting is one of the common defects in gear faults [1, 2]. At present, the mainstream gear fault diagnosis method is to analyze the vibration signal for gearbox fault diagnosis [3–6], and the obtained results cannot judge the type, size, shape, position and other detailed information of the gear fault. According to the standard requirements of "Gear Surface Bearing Capacity Test Method" (GB/T 14229-93), gear pitting area ratio is an important index to judge whether the gear is failure. The traditional measurement of gear pitting area rate mainly relies on manual visual inspection or microscope observation. This method has low detection efficiency, and the working time and efficiency are easily affected by human factors. Therefore, this study discusses how to accurately identify and measure the gear pitting rate on the basis of machine vision and deep learning (DL).

At present, two kinds of gear pitting detection methods exist. One uses vibration signals for diagnosing the pitting, and the other uses images for quantitatively detecting the pitting. Generally, the diagnosis [7] and prognosis

*Correspondence:
Yi Qin
qy_808@cqu.edu.cn
[1] State Key Laboratory of Mechanical Transmission, Chongqing University, Chongqing 400044, China
[2] College of Mechanical Engineering, Chongqing University, Chongqing 400044, China

[8] of gear pitting rely on vibration signals. However, the gear vibration signals collected from rotating machineries often have a huge amount of noise and are disturbed by the vibration signals of other components, making the quantitative detection of the degree of gear pitting difficult. Thus, gear pitting detection methods based on machine vision technology have been explored [9, 10]. Actually, the surface defect detection of industrial products is mainly based on machine vision and image processing [11, 12]. Traditional image processing methods for defect detection include threshold segmentation algorithm based on pixel statistics and edge detection algorithm based on background reconstruction [13, 14]. Although these methods have high segmentation accuracy on high-quality pitting images, they have low segmentation accuracy and poor generalization ability and robustness on pittings with different gray levels, textures, shapes, colors, and severity levels.

With the development of deep learning, deep neural networks are widely used in object detection, image classification, image segmentation and other fields. Deep neural networks with strong feature learning abilities have been widely used in the field of computer vision. Owing to the training of many annotated visual data, the features learned by deep neural networks contain a wealth of spatial and semantic information that can be applied well in object segmentation [15–17]. A new deep neural network is explored in this study to improve the segmentation precision and generalization ability of DL-based models and increase the accuracy of gear pitting measurement.

2 Related Work

Convolutional neural networks in deep learning technology [18] are widely used in various fields due to their powerful feature extraction capabilities, such as object detection, image classification, and image segmentation. The advantages of CNN are mainly in three aspects: (1) local receptive field (sparse connection): The original image is perceived through local convolution operation, and the feature image with global information is obtained with a greatly reduced number of weight parameters. (2) Parameter sharing: Due to the same convolution kernel parameters, the complexity of the input high-dimensional data can be reduced. (3) Multi-kernel convolution: Multi-layer convolution can be set in the convolution operation, and convolution of different convolution kernels can be set in each layer of convolution operation to extract different types of features. The FCN proposed by Long et al. [19] is a landmark network model in the field of image segmentation. Using the convolutional layer to replace the fully connected layer of CNNs, it is the first end-to-end fully convolutional network for pixel-level prediction. The advantage of FCN is that it can accept input images of any size, which makes feature extraction more efficient and avoids the problems of repeated calculation and space waste caused by the use of neighborhood. After the proposed network, the subsequent image segmentation algorithms are basically implemented based on this basic framework. For example, Ronneberger et al. [20] proposed the U-Net for the medical image segmentation processing task. U-Net adopts the encoder–decoder structure and splices the down-sampling convolution information with the up-sampling convolution information at the same layer, greatly improving the accuracy of segmentation. The SegNet network proposed by Badrinarayanan et al. [21] also inherited the encoder–decoder structure, but the decoder network used the pooling index of the maximum pooling layer to conduct nonlinear up-sampling. The DeepLab network proposed by Chen et al. [22] adopted a dilated convolution structure and optimized the boundary prediction through the fully connected conditional random field.

In recent years, attention models have been successfully introduced into the field of DL and widely applied to various fields, such as image processing and speech recognition [23–25]. After quickly scanning the global image, the human vision mechanism can apply different attentions on the image area. This method can improve the efficiency of the visual system in processing information by selecting key information from a large amount of information. The goal of the attention mechanism in DL is the same as the human visual mechanism. With the attention mechanism, the segmentation performance of U-Net can be also enhanced. In this paper, a novel type of U-Net is explored by splicing multi-scale features with attention modules, that is, multi-scale splicing attention U-Net (MSSA U-Net). The multi-scale splicing attention operation is essentially to fusing multi-scale features; thus, small objects (pittings) can be effectively recognized and segmented.

On the basis of MSSA U-Net, an integrated method for measuring the area ratio of gear pitting is proposed, and its flowchart is shown in Figure 1. The process of gear pitting detection generally includes the acquisition of pitting images, preprocessing of pitting images, segmentation of pitting images, and measurement of pitting area rate. First, a CCD camera and light source are installed on the gear contact fatigue testing machine to acquire pitting images. Secondly, the pitting image is preprocessed by geometric transformation, morphological operation, edge segmentation and other image processing technologies to obtain high-quality effective tooth surface images. Then, the pitting in the image is segmented by the trained MSSA U-Net. Finally, the proposed pitting area

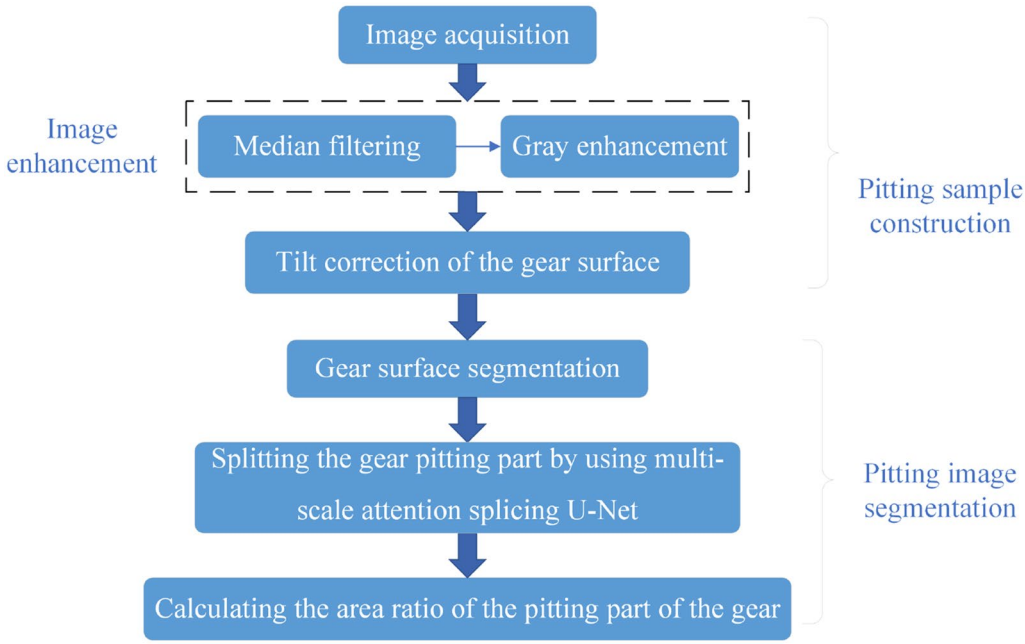

Figure 1 Flowchart of the proposed methodology for pittings detection

rate evaluation index is used to quantitatively evaluate the gear. The main contributions of this work are listed below.

To obtain rich characteristic information, including feature maps at different scales, an image splicing module is constructed to fuse the features from different convolutional layers.

Combining the image splicing module and an attention module, the MSSA U-Net is proposed to improve the segmentation precision of pittings.

With the comprehensive evaluation indexes, including four typical indexes, the performance and advantage of the proposed MSSA U-Net are verified on a gear pitting dataset and two public datasets.

3 Multi-Scale Splicing Attention U-Net

3.1 Preliminaries

U-Net is an improved image segmentation network based on FCN, which is a hot network in the field of medical image segmentation. A fully convolutional neural network with a U-shaped structure is constructed through skip connections and splicing operations, so it is called U-Net. The cross-entropy loss function it uses is written as

$$E = \sum_{x \in T} w(x) \log(p_{l(x)}(x)), \tag{1}$$

where T represents the pixel set of the real label, $p_{l(x)}(x)$ denotes the soft-max operation, and $w(x)$ is the weight parameter, and its calculation formula is as follows:

$$w(x) = w_c(x) + w_0 \exp\left(-\frac{[d_1(x) + d_2(x)]^2}{2\eta^2}\right), \tag{2}$$

where w_c is the weight to balance the proportion of categories, d_1 represents the distance from the nearest boundary, d_2 denotes the distance from the subnearest boundary, w_0 and η are constant values. In the subsequent training process, w_0 and η are set to 10 and 5 pixels, respectively [20].

U-Net mainly includes two parts: Encoder and decoder. The encoder is the down-sampling process, implemented by convolutional layers with max pooling operation. The decoder is the up-sampling process, implemented through deconvolution layers. Its overall structural framework is shown in Figure 2 [20]. Specifically, U-Net fuses the features of the same layer convolution and deconvolution stages through skip connections and Concat operations to ensure that more shallow semantic feature information is retained and the segmentation accuracy of the model is improved.

3.2 Construction of the Proposed Network

Conventional U-Net directly combines the feature map of the down-sampling layer with that of the up-sampling layer at the same stage. Evidently, the up-sampling

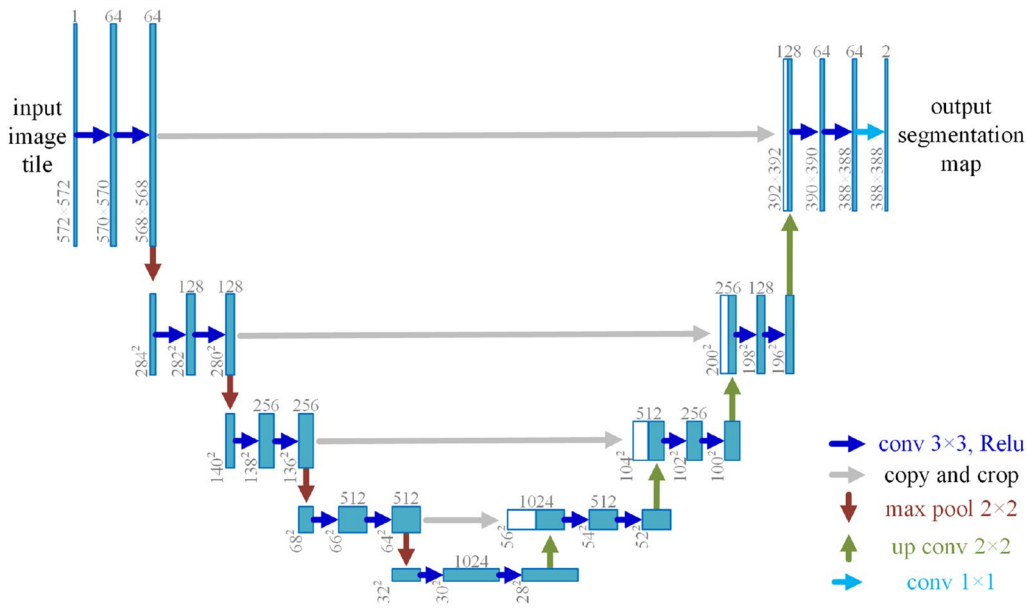

Figure 2 Structure of U-Net

operation of U-Net only uses the information of the previous layer to reconstruct the feature map, and the feature map of the down-sampling process only copies the feature map of the same layer, without considering that some important details contained in other layers may be lost, which may affect the quality of the recovered feature map [26]. To solve this problem, multi-scale supervision is realized by integrating the features of different convolutional layers. By introducing the attention mechanism and multi-scale splicing, the novel MSSA U-Net is developed, and its network structure is illustrated in Figure 3. The framework of MSSA U-Net also includes four down-sampling and four up-sampling processes. As the feature maps at different layers have different dimensions,

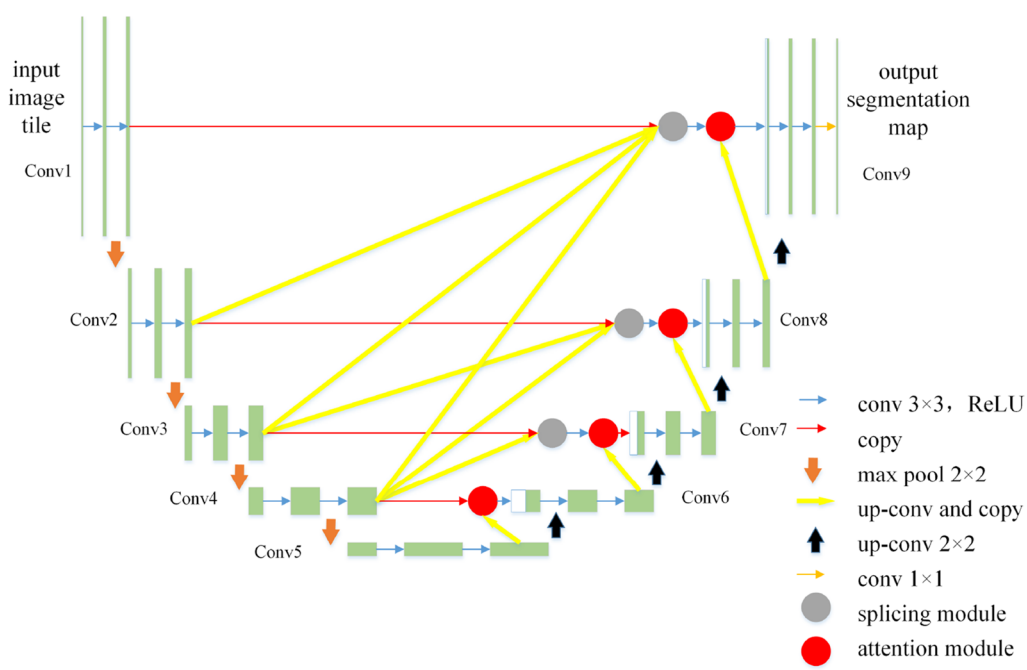

Figure 3 Structure of multi-scale splicing attention U-Net

the up-convolution operation should be used to concatenate the output image information of different convolutional layers into a higher-resolution feature map with more semantic characteristics and texture details. "up-conv 2×2" denotes the convolution with 2×2 kernel and a stride of 2, which can implement the upsampling operation. Then, the network uses the attention gate to select important information from the splicing feature map to improve the segmentation accuracy of the targets. As shown in Figure 3, the entire training process of the MSSA U-Net can be regarded as the encoding and decoding process in which the four down- and up-sampling operations are respectively regarded as encoding and decoding. The proposed network can combine the information of multi-layer feature maps, that is, achieve multi-scale critical feature fusion.

Unlike U-Net, an image splicing module is designed to splice the features from different convolutional layers in MSSA U-Net, as illustrated in Figure 4. To implement splicing, up-convolution should be applied first to adjust the feature maps of different dimensions to have the same dimension, as shown in Figure 3. n feature maps with F_0 channels are concatenated using Eq. (3):

$$X_0 = X_1 \parallel \ldots \parallel X_n, \tag{3}$$

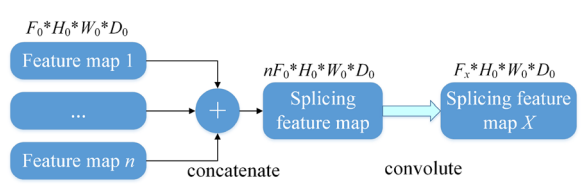

Figure 4 Image splicing module

where $\parallel$ denotes the splice operator.

Given that the splicing feature map has a different dimension from the up-sampling feature map, it requires dimensionality reduction. Convolution is performed on the splicing feature map for the two feature maps to have the same dimension. The convolution operation is defined as follows:

$$X = \sigma\left(W_{X_0}^{\mathrm{T}} X_0 + b_0\right), \tag{4}$$

where σ denotes the ReLU function. Then, the key features of the splicing feature map are selected via an attention module [23], which is illustrated in Figure 5. The attention module uses a feature map with high semantic feature (i.e., up-sampling feature map) to obtain a weight map (α^l) of the splicing feature map. As shown in Figure 5, splicing feature map X and up-sampling feature map Y are employed to construct a weight map, and then weighting processing is performed on X. According to Figure 5, α^l can be formulated as

$$\alpha^l = \varphi\left(F^{\mathrm{T}}\left(\sigma\left(W_X^{\mathrm{T}} X^l + W_Y^{\mathrm{T}} Y^l + b_W\right)\right) + b_F\right), \tag{5}$$

where $W_X^{\mathrm{T}} X^l$ and $W_Y^{\mathrm{T}} Y^l$ respectively represent the 1×1 convolution results of splicing feature map X and up-sampling feature map Y, F^{T} denotes a 1×1 convolution operation, b_W and b_F are the offset values, σ denotes the ReLU function, φ denotes the sigmoid function, and l indicates that the calculated feature map is located at layer l. Then, weight map α^l is multiplied by X, and weighted splicing feature map $\overline{X^l}$ is obtained as

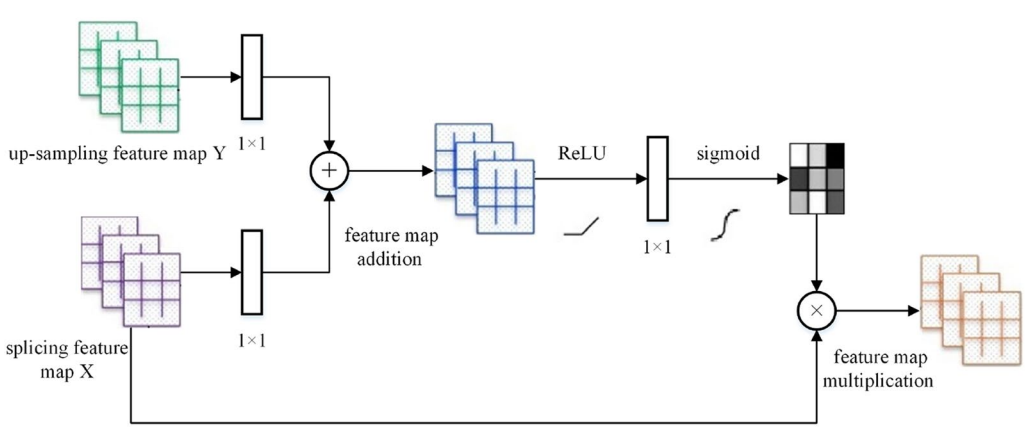

Figure 5 Attention module

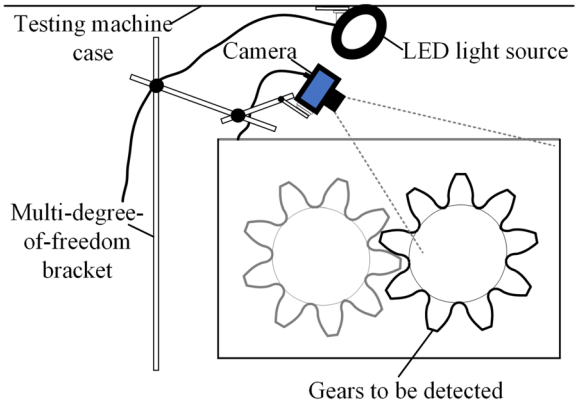

Figure 6 Schematic of the visual detection platform

$$\overline{X^l} = \alpha^l X^l. \tag{6}$$

From Eq. (6), the back propagation gradient of MSSA U-Net can be written as

$$\begin{aligned}\frac{\partial\left(\overline{X^l}\right)}{\partial\left(\varphi^{l-1}\right)} &= \frac{\partial\left(\alpha^l f\left(X^{l-1},\varphi^{l-1}\right)\right)}{\partial\left(\varphi^{l-1}\right)} \\ &= \alpha^l \frac{\partial\left(f\left(X^{l-1},\varphi^{l-1}\right)\right)}{\partial\left(\varphi^{l-1}\right)} + \frac{\partial\left(\alpha^l\right)}{\partial\left(\varphi^{l-1}\right)} X^{l-1}.\end{aligned} \tag{7}$$

In the traditional U-Net, low-resolution feature maps are mainly applied to recognize pittings, while high-resolution feature maps are mainly used to segment pittings. To detect irregular small pittings, MSSA U-Net uses more low-resolution key information at multiple scales to achieve object recognition, further improving the accuracy of pitting measurement.

4 Methodology for Gear Pitting Measurement

4.1 Pitting Dataset Construction and Preprocessing

First, we obtain gears with different severity levels of pitting by means of a gear contact fatigue testing machine. Secondly, the tooth surface image with pitting of each gear is collected by the designed gear pitting collection device. The gear pitting collection device is shown in

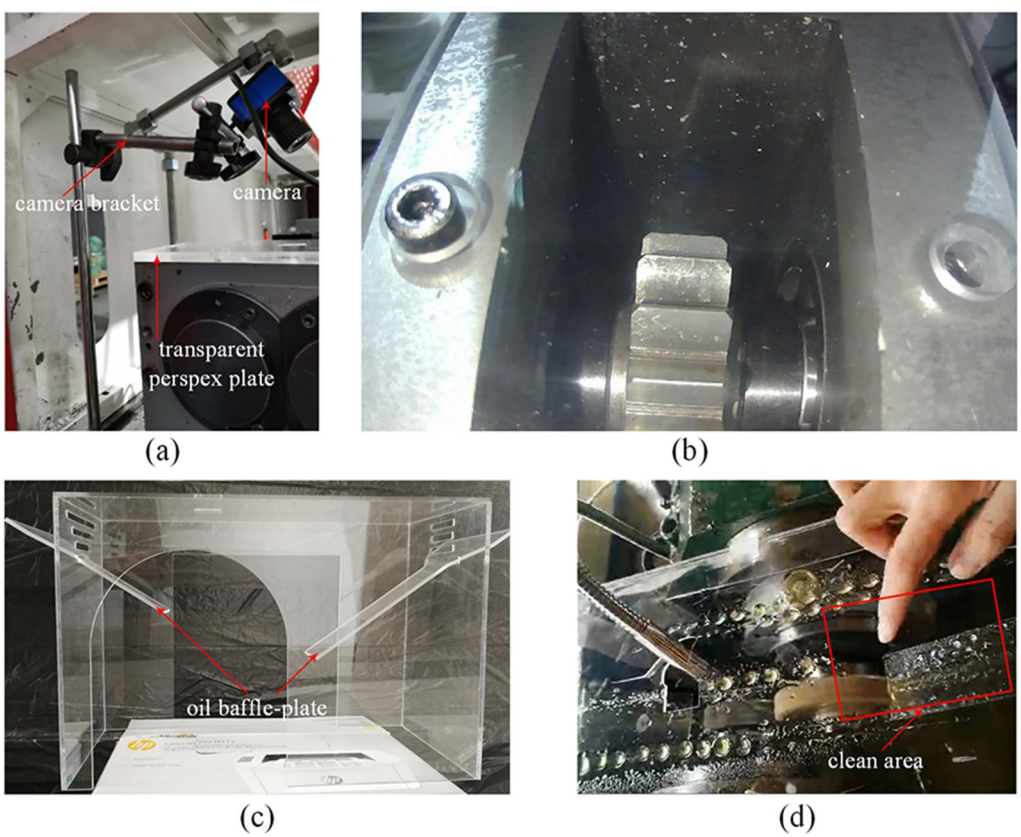

Figure 7 Visual detection platform for gear pitting: **a** visual detection device, **b** image acquisition, **c** case with oil baffle-plate, **d** effect of oil baffle

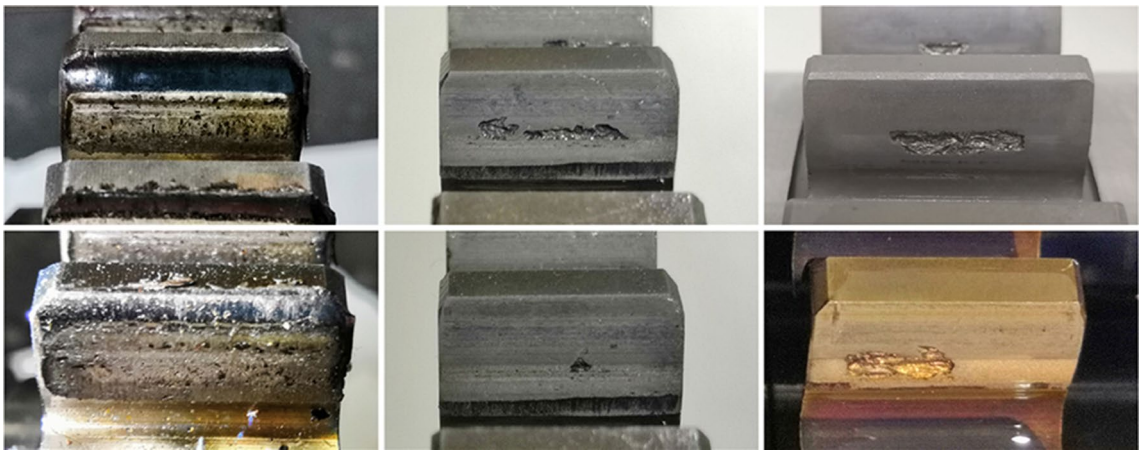

Figure 8 Partial presentation of the gear pitting dataset

Figures 6 and 7. The gear pitting collection device consists of a CCD camera (MER-131-210U3C), a circular LED light, and a set of adjustable fixed brackets. The collected original gear image is shown in Figure 8. Finally, a trainable pitting data set is produced through image preprocessing technology, and the severity of pitting in the data set is 1%−7%. According to a previous work [27], an appropriate lighting method was implemented for collecting clear pitting images.

During the test, the lubricating oil in the gearbox will splash as the gear rotates. This blocks the lens of the camera and prevents the gear from being photographed. Therefore, a detachable oil baffle mechanism is designed,

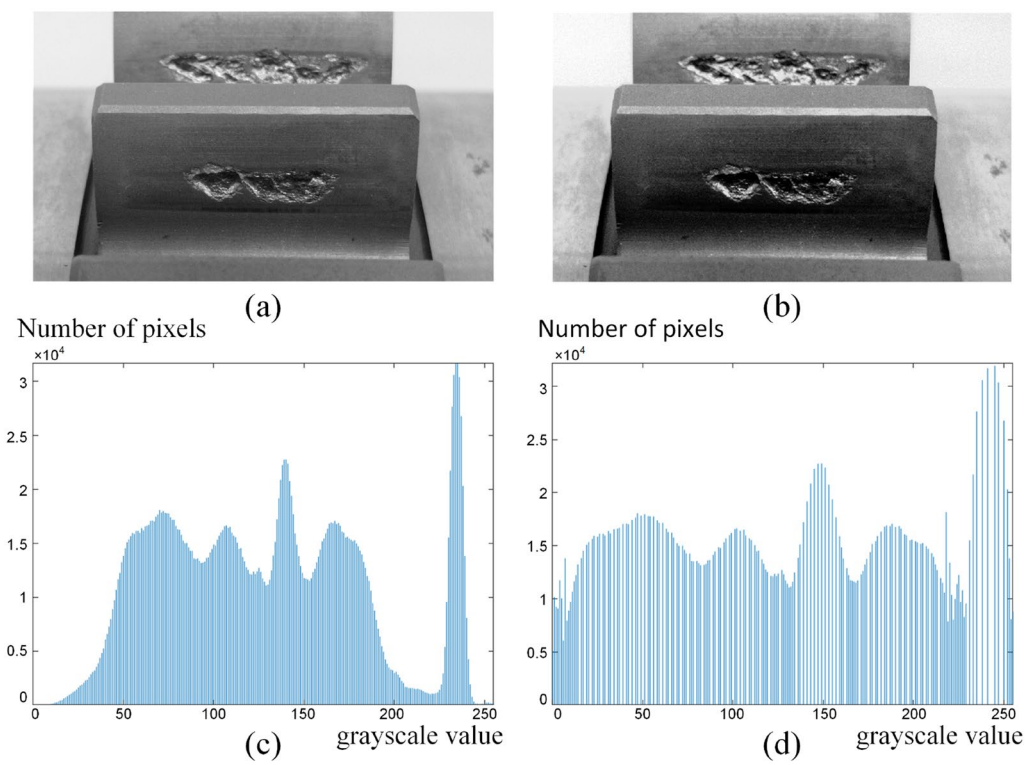

Figure 9 Tooth surface of the histogram equalization: **a** original image, **b** result image, **c** gray distribution of acquired image, **d** gray distribution of enhanced image

and the oil baffle plate is optimized through multiple tests, as shown in Figure 7(c) and (d).

To reduce the background interference, the distance and focal length between the camera and the single tooth surface were adjusted to obtain the image of the single tooth surface, as shown in Figure 8.

Due to the unstable indoor ambient light and the vibration of the test bench, noise will be introduced to the collected images. Therefore, the original gear image quality is improved by image preprocessing techniques. First, the original image is denoised using a median filter algorithm. Second, use histogram equalization to enhance the contrast of the image and make the edges of the gears clearer. The image after grayscale enhancement is shown in Figure 9. Figure 9(a) and (b) illustrate the originally acquired and enhanced images, respectively, and Figure 9(c) and (d) illustrate their corresponding gray distributions.

Figure 9(c) and (d) show that the gray value of the original image is relatively concentrated and changes greatly. The range of the gray value of the enhanced image is expanded, and its distribution is uniform, thereby improving the contrast of the image and highlighting the key information in the image.

Taking into account the impact of the self-vibration of the test bench on the acquisition device. It is also taken into account that in many different experiments, the shooting angles during the image acquisition process will be more or less different. Therefore, the image after image quality enhancement needs to be corrected for tooth surface inclination to improve the segmentation accuracy of the subsequent effective tooth surface. The tooth surface correction operation is completed by a powerful Radon transformation algorithm. The detailed implementation steps are as follows: (1) Use the edge detection operator to detect the horizontal line in the image to realize the enhancement of the gear edge. (2) Obtain the inclination angle of the gear tooth surface compared to the horizontal plane by calculating the Radon transformation of the image. Among them, the Radon transform is calculated as follows:

$$g_\theta(R) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)\delta(x\cos\theta + y\sin\theta - R)\mathrm{d}x\mathrm{d}y, \tag{8}$$

where $f(x,y)$ denotes a binary image matrix, called a density function. First, integrate f through the unit impulse function δ and the straight line in different directions $\left(x\cos\theta + y\sin\theta = R\right)$ to obtain the corresponding $g_\theta(R)$ after radon transformation, that is, the brightness value in different spaces. Then, find the local maxima of the radon transformation $g_\theta(R)$, and the positions of these maxima are the positions of the straight lines in the original image (Figure 10(a)) (the direction of the straight line is θ). Finally, the inclination correction of the gear is completed by the rotation operation, as shown in Figure 10(b). Figure 8 indicates a satisfactory result. Tilt correction helped obtain an accurate effective tooth surface area, thereby improving the precision of the pitting area ratio calculation.

4.2 Tooth Surface and Pitting Segmentation

Next, before segmenting pittings, the effective tooth surface is required to be segmented out. The effective tooth surface is segmented using traditional image processing methods. Firstly, the k-means clustering algorithm is used to cluster the image to determine the gray threshold range of the gear surface. Secondly, use the Roberts differential operator to detect the edge, and get the edge information of the tooth surface that needs to be segmented. Since the gear tooth surface area to be segmented in the image has obvious characteristics and is completely different from the texture features of the background, the neural network method is not used to segment the effective tooth surface. During the solution process, the horizontal difference and vertical difference of the image are approximated as:

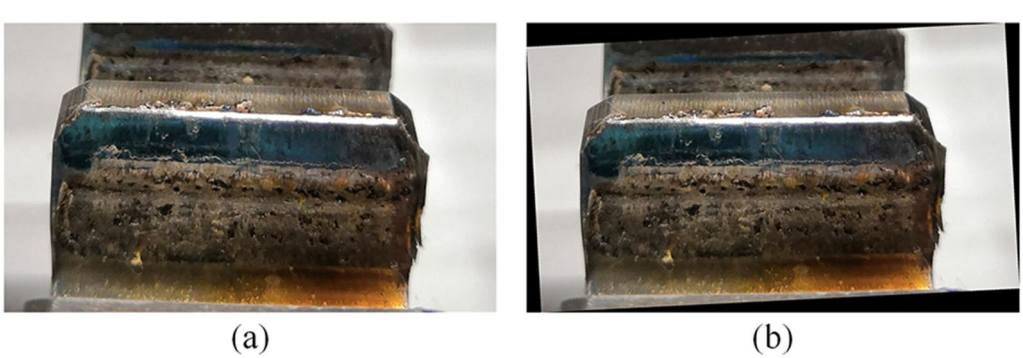

Figure 10 Comparison of gear tooth surface tilt correction: **a** original image, **b** result image

Figure 11 The result display of the segmentation process of the effective working tooth surface: **a** preliminary segmentation result, **b** binary image, **c** segmentation result

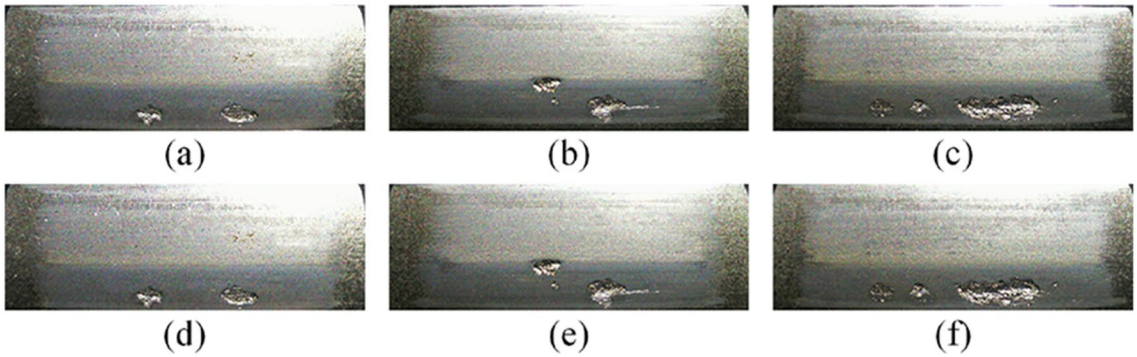

Figure 12 Different effective working tooth surfaces: **a**–**c** segmented effective tooth surfaces, **d**–**f** ground truths of effective tooth surfaces

$$\nabla f = (f(x,y) - f(x-1,y), f(x,y) - f(x,y-1)), \quad (9)$$

where $f(x,y)$ denotes the pixel value of point (x,y) in the image matrix.

For the tooth surface segmentation of the pitting image after image enhancement and tilt correction, the preliminary segmentation results using the above traditional segmentation method are shown in Figure 11(a). It can be seen from the figure that the effective tooth surface area is roughly divided, and there are problems such as discontinuous tooth surface edges and many edge fractures. Therefore, it is necessary to perform subsequent refinement operations on this result in order to segment the complete effective tooth surface area. First, binarize the pre-segmented image to obtain a binary image. Then, dilate and corrode the binary image to obtain a relatively complete connected area, as shown in Figure 11(b). Finally, the original image is segmented according to the coordinate information of its maximum connected area to obtain an accurate image of the effective tooth surface area, as shown in Figure 11(c). The display results of partially segmented effective tooth surfaces in the data set are shown in Figure 12(a)–(c).

5 Experimental Results

5.1 Operating Environment and Evaluation Indicators

Python 3.6.8, TensorFlow-gpu 1.7.0, Keras 2.3.1, and other common packages (numpy 1.16.4, matplotlib 2.2.2, etc.) were used to train and test the MSSA U-Net. The training process was implemented in a computer server with the following specifications: an Nvidia GeForce GTX-1080Ti GPU and an Intel (R) Xeon (R) E5-2687W v3 CPU. All neural networks have the same hyper parameters, which are listed in Table 1.

Table 1 Hyper parameters of all used neural networks

Hyper parameter	Value
Learning rate	0.0025
Optimization algorithm	SGD
Multiple of learning rate change	0.1
Step of learning rate change	20000
Rotation_range	0.2
Height_shift_range	0.05
Width_shift_range	0.05
Zoom_range	0.05
Shear_range	0.05
Steps_per_epoch	1000
Epochs	400
Batch size	2

In order to better evaluate the accuracy of the proposed model, several widely used classic evaluation indicators

are introduced, including precision P, recall rate R [28], and intersection-over-union ratio IoU [29]. Since our proposed model is used for the detection and measurement of gear pitting. Therefore, in order to quantitatively evaluate the detection accuracy of gear pitting, it is comprehensively evaluated by calculating the relative error Re between the detected pitting area rate and the actual pitting area rate. In addition, the classic evaluation index harmonic mean F_1 is additionally introduced, so that the performance of the proposed method can be comprehensively evaluated:

$$P = \frac{T_P}{T_P + F_P}, R = \frac{T_P}{T_P + F_N}, F_1 = \frac{2 \times P \times R}{P + R}, \quad (10)$$

where T_P is the positive sample predicted by the model as a positive class. T_N is the negative sample predicted by the model as a negative class. F_P is the negative sample predicted by the model as a positive class. F_N is the positive sample predicted by the model as a negative class.

The mean intersection-over-union ratio (MIoU) is one of the important indicators for evaluating image semantic segmentation networks. It is obtained by calculating the IoU of each category [30]. The specific calculation process is as follows:

$$MIoU = \frac{1}{k+1} \sum_{i=0}^{k} \frac{T_P}{F_N + F_P + T_P}, \quad (11)$$

where $k+1$ represents the number of categories (counting the number of categories of the background).

5.2 Dataset Introduction

In order to realize the detection task of gear pitting based on deep learning technology, a data set of gear pitting with different severity pitting was constructed. This data set can meet the requirements of model training and testing. In addition, two public datasets are used to further verify the superiority of the proposed model MSSA U-Net. The details of the datasets are shown in Table 2.

Gear pitting dataset: In the contrast experiment, all 2200 sample images were divided into a training set with 2000 images and a test set with 200 images. Then, all the gear pitting areas of the training set were labeled. Labelme, an open-source image annotation tool developed by MIT Computer Science and Artificial Intelligence Laboratory (CSAIL), was used to generate the ground truth mask. With the trained MSSA U-Net, the test set was used for evaluating the proposed methodology.

CrackForest dataset [31]: This dataset has a similar defect type to the gear pitting dataset, which is a road crack dataset. There are 159 images in this dataset, 6 of which are road images without defects. Due to the small number of data sets, data expansion methods (random flip, mean subtraction) are used to expand the data set. After expansion, 1500 images are obtained, of which 120 are used for training and 300 are used for testing.

Voc2012 dataset [32]: This dataset can be used for classification, detection and segmentation. We use the enhanced Voc2012 dataset, which has a total of 20 object categories. 10582 images are used as training set, 1449 images are used as validation set, and 1456 images are used as test set.

In the experiments, all the images were resized to 512×512 for fair comparison. Especially for the gear pitting dataset, the new resolution can improve the measurement efficiency of gear pitting. In addition, the minimum and maximum sizes of the segmentation objects in three datasets are listed in Table 2.

5.3 Comprehensive Evaluation Indexes

The accurate segmentation of the effective working tooth surface is the key to calculate the gear pitting area rate. The image is cropped by the crop function to obtain the ground truth of the effective tooth surface image. The ground truths of the segmented effective tooth surfaces shown in Figure 12(a)–(c) are illustrated in Figure 12(d)–(f). If A_t denotes the predicted effective working tooth surface area and A_{tg} denotes the area of the effective tooth surface ground truth, then the segmentation accuracy (P_t) of the effective tooth surface is defined as follows:

$$P_t = 1 - \frac{|A_t - A_{tg}|}{A_t} \times 100\%. \quad (12)$$

Table 2 Details of three datasets

Name	Number of images	Number of testing images	Resolution	Minimum object size	Maximum object size
Gear pitting	2200	200	1280 × 1024	15 × 15 pixels	1213 × 235 pixels
Crackforest	1500	300	480 × 320	57 × 37 pixels	480 × 246 pixels
Pascal VOC	13487	1456	500 × 375	54 × 43 pixels	489 × 334 pixels

The P_t value of each image in the gear pitting data set can be calculated by using Eq. (12), and then added and averaged to obtain the segmentation accuracy of the effective tooth surface in this paper, 96.3%. Next, the method MSSA U-Net proposed in this paper is used to segment the pitting area.

According to the segmented tooth surface and pitting, effective working tooth surface area A_t and pitting area A_p are computed by counting the number of pixels in the segmented regions, and then the calculation formula of gear pitting area ratio (AR) is as follows:

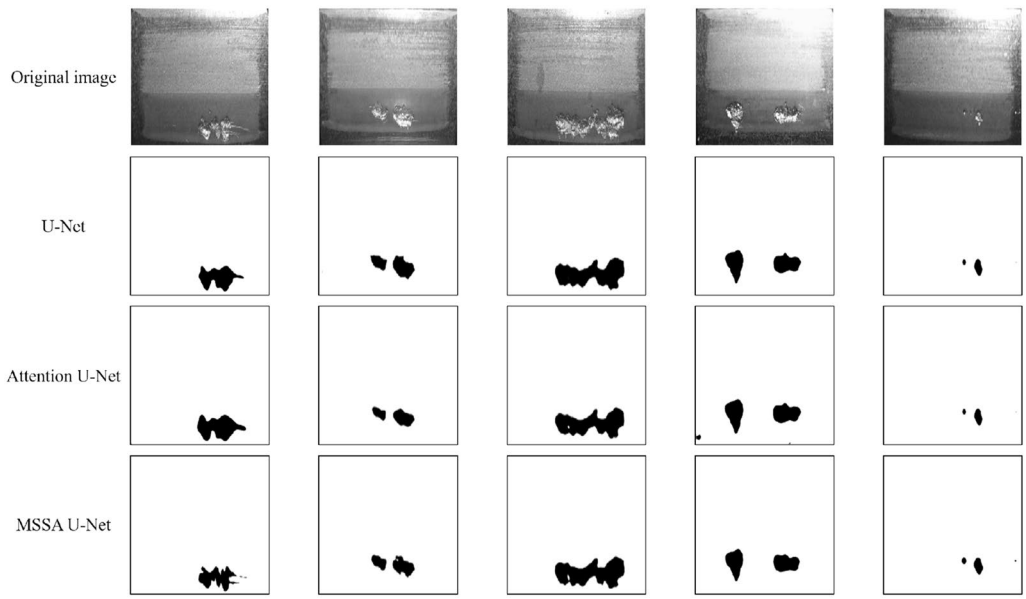

Figure 13 Pitting segmentation results of different neural networks

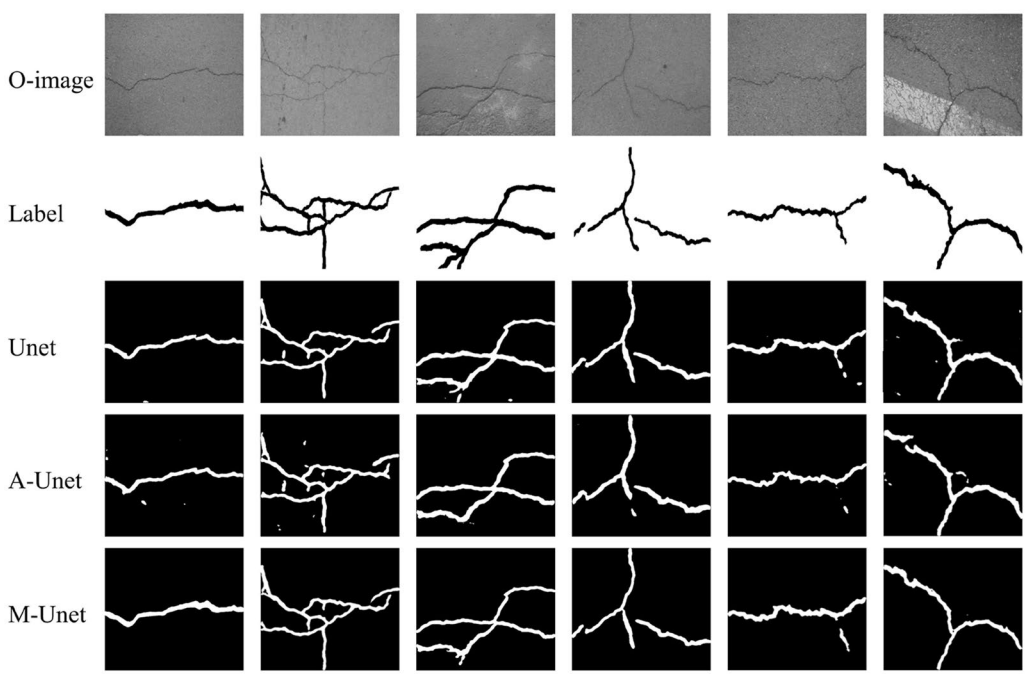

Figure 14 Typical segmentation results of the "Crackforest" dataset obtained by various neural networks

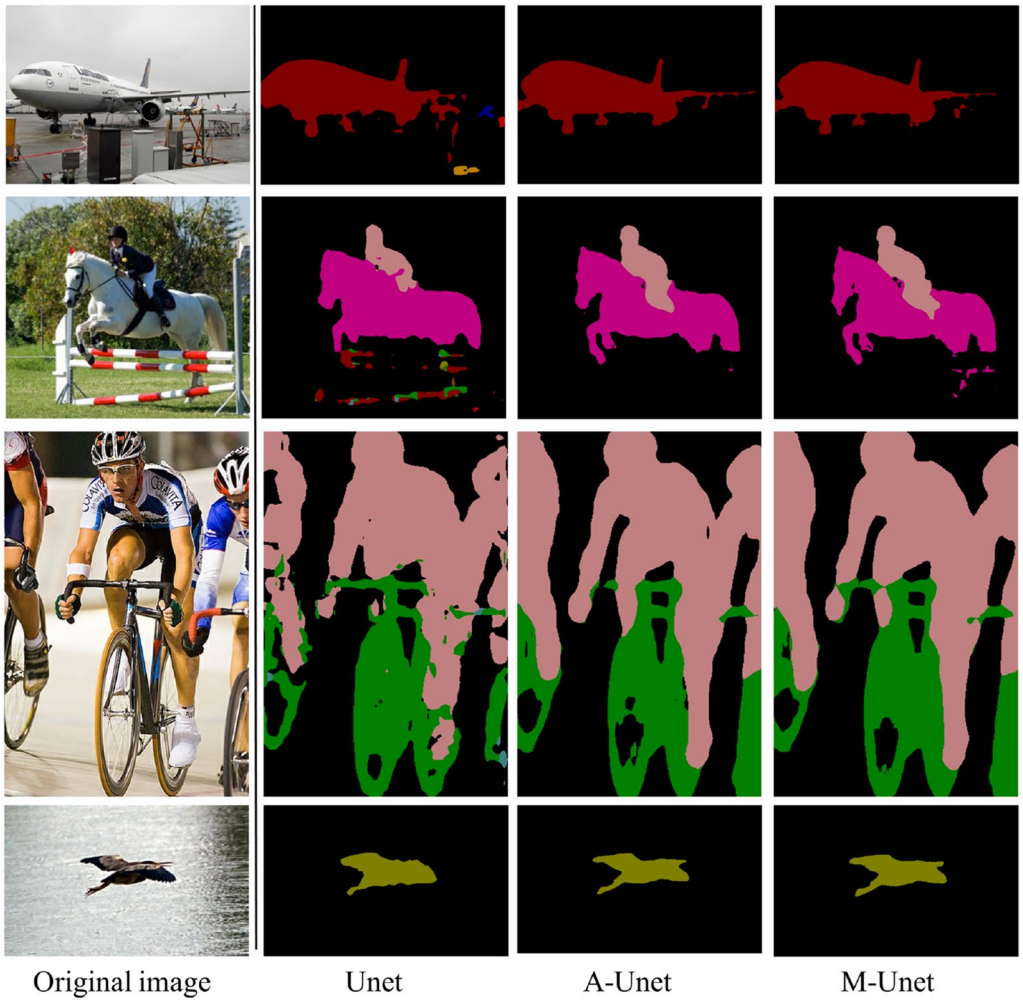

Figure 15 Several examples of segmentation results obtained by different neural networks

$$AR = \frac{A_p}{A_t} \times 100\%. \tag{13}$$

By calculating the intermediate value of the ratio of the predicted pitting area to the actual pitting area (Eq. (13)), this paper proposes a relative error (*Re*) evaluation index. This index can not only obtain the pitting area ratio, but also evaluate the accuracy of the proposed method. The relative error is calculated as follows:

$$Re = \left| \frac{AR_p - AR_a}{AR_a} \right| \times 100\%, \tag{14}$$

where AR_p denotes the detected pitting area ratio. AR_a represents the actual pitting area ratio.

5.4 Results Visualization

For comparison, several U-types of neural networks, such as U-Net and attention U-Net, were trained on the same training sample set, and then the test sample set was used for comprehensive evaluation. The segmentation results obtained by three neural networks are illustrated in Figures 13, 14, and 15. To further verify the superiority of our proposed model, two open datasets were used for comparison. Figure 13 shows that MSSA U-Net could capture more minor details on the irregular pittings and has a better segmentation performance than the other

Table 3 Comparison results of different network models

Model	*Re*	*P*	*R*	MIoU
U-Net	7.80	86.93	90.56	79.80
Attention U-Net	10.55	87.84	83.11	74.41
MSSA U-Net	**6.23**	**87.97**	**89.82**	**80.17**

Bold values indicate that the detection accuracy of the proposed method is significantly better than that of other methods

Table 4 Training and running times of three models for the gear pitting dataset

Model	Training time (h)	Running time (s)
U-Net	12.10	2.95
Attention U-Net	13.58	3.42
MSSA U-Net	15.44	3.71

U-Nets. Attention U-Net directly employs the attention module in the concatenation of one layer but does not adequately mine the key semantic information of the feature maps at different scales. Consequently, some minor details of pitting cannot be correctly identified. Owing to the lack of multi-scale semantic feature fusion and attention mechanism, the segmentation performance of the classical U-Net is also worse than that of MSSA U-Net.

The comprehensive evaluation indexes obtained by the three U-nets are listed in Table 3. Evidently, MSSA U-Net has the highest measurement accuracy among the U-nets. The proposed MSSA U-Net has a higher precision rate than U-Net, but its recall rate is slightly lower, mainly because the precision and recall rates are mutually contradictory, and the improvement of one will lead to the relative decrease of the other. In this research, we focused on improving the precision rate while ensuring a high recall rate. With MSSA U-Net, the *Re* index decreased to 6.23%, that is, the measurement accuracy reached 93.77%, which can meet the gear pitting measurement requirement. Moreover, MSSA U-Net has a stronger robustness in pitting segmentation even though the acquired gear teeth images suffer from oil pollution. Only 28 pitting images in the test set has a relative error rate of more than 10%, of which the largest is 16.22%. Table 4 shows that all the four indexes obtained by MSSA U-Net are superior to those obtained by attention U-Net. Similarly, the traditional image segmentation method (threshold segmentation) was used to measure the pitting area ratio; the obtained *Re* is as high as 31.79%, and its precision is much lower than that of the proposed methodology, further demonstrating the superiority of MSSA U-Net over the conventional image segmentation method. According to the comparison of the comprehensive evaluation indexes, the proposed vision measurement methodology based on MSSA U-Net has a high accuracy and can be effectively applied in engineering practice.

Table 5 Comprehensive evaluation indexes of "Crackforest" dataset segmentation obtained by several networks

Model	*P*	*R*	MIoU	F_1
U-Net	90.37	89.74	85.01	90.05
Attention U-Net	88.16	86.10	82.43	87.12
MSSA U-Net	**93.61**	**91.02**	**88.75**	**92.29**

Bold values indicate that the detection accuracy of the proposed method is significantly better than that of other methods

Table 6 Comprehensive evaluation indexes obtained by three networks for the "VOC2012" dataset

Model	*P*	*R*	MIoU	F_1
U-Net	80.81	79.77	70.85	78.38
Attention U-Net	77.48	76.49	67.72	75.51
MSSA U-Net	**82.54**	**81.80**	**73.43**	**80.40**

Bold values indicate that the detection accuracy of the proposed method is significantly better than that of other methods

The training time of each model is obtained and listed in Table 4. After training the three models, the running time of each model for processing a test image is also listed in Table 4. Given its highly complex network structure, the calculated amount of the proposed MSSA U-Net is larger than those of U-Net and Attention U-Net.

For Attention U-Net and MSSA U-Net, the obtained typical segmentation results on the gear pitting dataset are shown in Figure 14. This figure shows that the segmentation result is very near the real objects. Then, the comprehensive evaluation indexes of various neural networks are calculated and listed in Table 5. Table 5 indicates that the proposed MSSA U-Net has a higher segmentation ability than U-Net and Attention U-Net.

For the "VOC2012" dataset, U-Net, attention U-Net, and MSSA U-Net were trained under the same hardware and software. Several examples of segmentation results obtained by different neural networks are shown in Figure 15. Then, the comprehensive evaluation indexes of the three networks were calculated, as listed in Table 6. This table shows that the MIoU obtained by MSSA U-Net is approximately 2.6% higher than that obtained by U-Net, and the values of the four indexes obtained by MSSA U-Net are significantly higher than those obtained by Attention U-Net. The comparative results further demonstrate the superiority of the proposed MSSA U-Net over the classical image segmentation network.

6 Conclusions

This study addresses the development of U-Net neural networks and successfully constructed a MSSA U-Net with a high segmentation performance. On the basis of MSSA U-Net and computer vision, a new gear pitting measurement methodology is proposed to accurately

measure the area ratio of gear pitting. Particularly, the proposed method can be applied in recognizing tiny irregular pittings. The main contributions of this work are as follows:

(1) A new image splicing module was first designed for fusing multi-scale semantic features.
(2) Combining the image splicing and attention modules, the MSSA U-Net model was built to improve the accuracy of pitting segmentation.
(3) Using comprehensive evaluation indexes, the comparative results illustrate that the proposed methodology has a higher segmentation performance than U-Net, attention U-Net, and traditional image segmentation methods.
(4) The proposed method has great potential in the quantitative evaluation of the gear pitting degree. On the basis of the proposed method, we will develop an on-line visual measurement system in future work.

Acknowledgements
Not applicable.

Author Contributions
YQ was in charge of the conceptualization and the methodology; DX was in charge of formal analysis and writing the draft; WC was in charge of the software and the investigation; YW was in charge of validation. All authors read and approved the final manuscript.

Authors' Information
Yi Qin received his B.Eng. and Ph.D. degrees in mechanical engineering from *Chongqing University, Chongqing, China*, in 2004 and 2008, respectively. Since January 2009, he has been with *Chongqing University*, where he is currently a professor of *College of Mechanical and Vehicle Engineering*. He has published over 120 papers and holds 14 invention patents. His current research interests include fault prognosis, artificial intelligence, digital twin, and visual detection.
Dejun Xi received her Master's degree in mechanical engineering in 2018 from *Northeast Agricultural University, Harbin, China*. She is currently pursuing a doctoral degree in mechanical engineering at *Chongqing University*, and her research interests are in intelligent manufacturing, pattern recognition, and image processing.
Weiwei Chen was born in 1994, received his Master's degree in mechanical engineering in 2020 from *Chongqing University, Chongqing, China*. His research interest includes machine vision.
Yi Wang received his B. Eng. degree from *Southwest Jiaotong University, Chengdu, China*, in 2011 and his Ph.D. degree from *Xi'an Jiaotong University, Xi'an, China*, in 2017. From August 2016 to February 2017, he was a visiting scholar in *City University of Hong Kong, Hong Kong, China*. Since July 2017, he has been a lecturer in *Chongqing University, Chongqing, China*. His current research interests include mechanical signal processing, weak signal detection, rotating machinery fault diagnosis under speed variation conditions, manifold learning, and deep learning.

Funding
Supported by National Natural Science Foundation of China (Grant Nos. 62033001 and 52175075), and Chongqing Municipal Graduate Scientific Research and Innovation Foundation of China (Grant No. CYB21010).

Competing Interests
The authors declare no competing financial interests.

Received: 15 January 2021 Revised: 8 March 2023 Accepted: 9 March 2023
Published online: 07 April 2023

References
[1] M Amarnath, S Lee. Assessment of surface contact fatigue failure in a spur geared system based on the tribological and vibration parameter analysis. *Measurement*, 2015, 76: 32-44.
[2] Y Qin, S Xiang, Y Chai, et al. Macroscopic-microscopic attention in LSTM networks based on fusion features for gear remaining life prediction. *IEEE Transactions on Industrial Electronics*, 2020, 67(12): 10865-10875.
[3] F Shen, C Chen, J Xu, et al. A fast multi-tasking solution: NMF-theoretic co-clustering for gear fault diagnosis under variable working conditions. *Chinese Journal of Mechanical Engineering*, 2020, 33(1). https://doi.org/10.1186/s10033-020-00437-3
[4] R Chen, X Huang, L Yang, et al. Intelligent fault diagnosis method of planetary gearboxes based on convolution neural network and discrete wavelet transform. *Computers in Industry*, 2019, 106: 48-59.
[5] Y Qin, Y Mao, B Tang, et al. M-band flexible wavelet transform and its application to the fault diagnosis of planetary gear transmission systems. *Mechanical Systems and Signal Processing*, 2019, 134: 106298.
[6] S Xiang, Y Qin, C Zhu, et al. Long short-term memory neural network with weight amplification and its application into gear remaining useful life prediction. *Engineering Applications of Artificial Intelligence*, 2020, 91: 103587.
[7] X Wang, Y Qin, A Zhang. An intelligent fault diagnosis approach for planetary gearboxes based on deep belief networks and uniformed features. *Journal of Intelligent & Fuzzy Systems*, 2018, 34: 3619-3634.
[8] Z Pu, D Cabrera, Y Bai, et al. A one-class generative adversarial detection framework for multifunctional fault diagnoses. *IEEE Transactions on Industrial Electronics*, 2021, 69(8): 1-11.
[9] D Xi, Y Qin, S Wang. YDRSNet: an integrated Yolov5- Deeplabv3+ real-time segmentation network for gear pitting measurement. *Journal of Intelligent Manufacturing*, 2021: 1-15.
[10] Y Qin, Z Wang, and D Xi. Tree CycleGAN with maximum di-versity loss for image augmentation and its application into gear pitting detection. *Applied Soft Computing*, 2022, 114: 108130.
[11] T Niu, B Li, W Li, et al. Positive-sample-based surface defect detection using memory-augmented adversarial autoencoders. *IEEE/ASME Transactions on Mechatronics*, 2022, 27(1): 46-57.
[12] J Wang, K Song, D Zhang, et al. Collaborative learning attention network based on RGB image and depth image for surface defect inspection of no-service rail. *IEEE/ASME Transactions on Mechatronics*, 2022.
[13] W Zhang, X Wang, W You, et al. RESLS: region and edge synergetic level set framework for image segmentation. *IEEE Transactions Image Process*, 2020, 29: 57-71.
[14] L Wang, L Xu, J Yu, et al. Context-aware edge similarity segmentation algorithm of time series. *Cluster Computing*, 2016, 19: 1421-1436.
[15] Z Zhu, P Luo, X Wang, et al. Deep learning identity-preserving face space. *IEEE International Conference on Computer Vision*, Sydney, Australia, December 3-6, 2013: 113-120.
[16] Y Sun, X Wang, X Tang. Deep learning face representation from predicting 10,000 classes. *IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, June 23-28, 2014: 1891-1898.
[17] Z Jin, J Yang, Z Hu, et al. Face recognition based on the uncorrelated discriminant transformation. *Pattern Recognition*, 2001, 34: 1405-1416.
[18] A Krizhevsky, I Sutskever, G Hinton. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 2017, 60: 84-90.
[19] E Shelhamer, J Long, T Darrell. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, 39: 640-651.

[20] O Ronneberger, P Fischer, T Brox. U-Net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention*, 2015, 9351: 234-241.
[21] V Badrinarayanan, A Kendall, R Cipolla. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, 39: 2481-2495.
[22] L Chen, G Papandreou, I Kokkinos, et al. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018, 40: 834-848.
[23] O Oktay, J Schlemper, L Folgoc, et al. Attention U-Net: Learning where to look for the pancreas. *IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, USA, June 19-21, 2018: 1804.03999.
[24] S Song, C Lan, J Xing, et al. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. *Proceedings of the AAAI Conference on Artificial Intelligence*, San Francisco, California USA, February 4-9, 2017, 31(1).
[25] G Hassan, A Hassanien, N Elbendary, et al. Blood vessel segmentation approach for extracting the vasculature on retinal fundus images using particle swarm optimization. *International Computer Engineering Conference*, Giza, Egypt, December 29-30, 2015: 290-296.
[26] C Deng, M Wang, L Liu, et al. Extended feature pyramid network for small object detection. *IEEE Transactions on Multimedia*, 2022, 24: 1968-1979.
[27] D Xi, Y Qin, Y Wang. Vision measurement of gear pitting under different scenes by deep mask R-CNN. *Sensors*, 2020, 20(15): 4298.
[28] C Goutte, E Gaussier. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. *Advances in Information Retrieval*, 2005, 3408: 345-359.
[29] F Ahmed, D Tarlow, D Batra. Optimizing expected intersection-over-union with candidate-constrained CRFs. *IEEE International Conference on Computer Vision*, Santiago, Chile, USA, December 7-13, 2015: 1850-1858.
[30] L Hamers, Y Hemeryck, G Herweyers, et al. Similarity measures in scientometric research - the jaccard index versus salton cosine formula. *Information Processing & Management*, 1989, 25: 315-318.
[31] Y Shi, L Cui, Z Qi, et al. Automatic road crack detection using random structured forests. *IEEE Transactions on Intelligent Transportation Systems*, 2016, 17(12): 3434-3445.
[32] M Everingham, S M A Eslami, L Van Gool, et al. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 2015, 111: 98-136.