scientific reports

Check for updates

OPEN

Research on the performance of the SegFormer model with fusion of edge feature extraction for metal corrosion detection

Bingnan Yan, Conghui Wang & Xiaolong Hao✉

Addressing the challenge that existing deep learning models face in accurately segmenting metal corrosion boundaries and small corrosion areas. In this paper, a SegFormer metal corrosion detection method based on parallel extraction of edge features is proposed. Firstly, to solve the boundary ambiguity problem of metal corrosion images, an edge-feature extraction module (EEM) is introduced to construct a spatial branch of the network to assist the model in extracting shallow details and edge information from the images. Secondly, to mitigate the loss of target feature information during the reconstruction of the decoder, this paper adopts the gradual upsampling decoding layer design. It introduces the feature fusion module (FFM) to achieve hierarchical and progressive feature fusion, thereby enhancing the detection of small corroded areas. Experimental results show that the proposed method outperforms other semantic segmentation models achieving an accuracy of 86.56% on the public metal surface corrosion image dataset and reaching a mean intersection over union (mIoU) of 91.41% on the BSData defect dataset. On the Self-built tubing corrosion pit image dataset, the model utilizes only 3.60 MB of parameters to achieve an accuracy of 96.52%, confirming the effectiveness and performance advantages of the proposed method in practical applications.

Keywords Image semantic segmentation, Metal surface corrosion image, Tubing corrosion pit image, Edge-feature extraction

Research indicates that annual economic losses caused by corrosion accounts for approximately 2% to 4% of the national gross domestic product (GDP), with around 10% of the steel produced globally being consumed by corrosion[1]. Due to the extensive and diverse corrosion phenomena, its occurrence and related costs are difficult to completely eliminate. However, most studies suggest that effective corrosion management practices could save 25% to 30% of annual corrosion costs[2]. In this context, the detection and protection of metal corrosion are particularly important, especially in the evaluation of material service status, failure analysis and remaining service life.

Traditional corrosion detection methods, such as manual inspection or detection based on traditional image processing, often rely on expert experience, are inefficient and susceptible to subjective factors. With the development of computer vision technology, the detection method based on a deep learning model has significantly enhanced the intelligence of metal surface corrosion detection. The main idea of this method is to use a deep learning model to semantically segment the pixels of a metal surface corrosion image, extract the texture, shape, and other features of the corrosion area through model training, and quantify the corrosion degree according to the corrosion area[3]. At present, metal surface corrosion image detection algorithms based on deep learning can be divided into two categories. One is to extract corrosion features from the image through image semantic segmentation based on CNN model architecture. The other is the metal surface corrosion detection method based on the Transformer model. The image semantic segmentation method based on CNN model architecture has outstanding performance in metal surface defect detection[4]. However, due to the limited receptive field of convolutional neural networks applied in extracting pixel features, the lack of comprehensive understanding of context information leads to the insufficient accuracy of image semantic segmentation, which affects the accuracy of metal surface corrosion detection. The Transformer model has the advantage of global perception. In the image semantic segmentation task, it can simultaneously consider the pixels in each position of the image, capture the long-distance dependency between pixels, and achieve an overall perception and high-level understanding of the entire image, thus compensating to some extent the shortcomings of the

School of Electronic Engineering, Xi'an Shiyou University, Xi'an 710065, China. ✉email: haoxl315024@163.com

CNN model in dealing with long-distance dependency[5–7]. However, although Transformer has demonstrated strong performance in image segmentation tasks, traditional Transformer models still face many challenges when processing high-resolution images due to heavy computation and low efficiency. To solve these problems, SegFormer was created. In recent years, many scholars have tried to apply the advantages of SegFormer to metal surface corrosion detection, achieving promising results[8–10]. SegFormer is a semantic segmentation model combining the Transformer architecture with an efficient decoding structure, which can make full use of the advantages of Transformers in capturing long-range dependencies and global context information, while avoiding the computational and efficiency bottlenecks of traditional Transformer models. Through adaptive feature extraction and cross-level feature fusion, SegFormer can achieve accurate segmentation of corrosion regions in complex environments, effectively improving the robustness and accuracy of the model.

However, although SegFormer demonstrates strong performance in metal surface corrosion detection, it still faces some challenges, especially when dealing with blurred boundaries and irregular shapes of the corrosion areas. Metal surface corrosion images often have boundary blurring, irregular shapes and other issues, which lead to detail loss, texture blurring and edge segmentation errors when deep learning models are used for pixel segmentation[11,12]. To solve the above problems, and to improve the semantic segmentation accuracy of image edge pixels, this paper proposed a SegFormer model based on parallel extraction of edge features. The main contributions of this paper include three aspects:

(1) An edge-feature extraction module (EEM) was proposed. Sobel operator, multi-scale convolution, and Laplacian operator were used to construct a parallel structure to extract the boundary features of the corrosion region and obtain the shallow Edge information of the corrosion region. The extra branches of the network constructed by edge operators improve the model's accuracy for boundary segmentation of irregular metal surface corrosion.
(2) Considering the limited accuracy of the model due to the simple structure of the SegFormer decoder, this paper adopted the design of a gradually up-sampling decoding layer, proposed the feature fusion module (FFM) to achieve hierarchical feature fusion, and gradually up-sampling to recover the details of the corroded image. Through skip connection, the low-level and high-level features are fused to integrate multi-scale information and improve the robustness and accuracy of model segmentation.
(3) To verify the effectiveness of the proposed method, this paper established a dataset of tubing corrosion pit images to verify the performance of the SegFormer model based on parallel edge-feature extraction in the actual tubing corrosion pit image detection task, providing a valuable reference for corrosion detection in industrial applications.

Related works

Corrosion detection

The detection of corrosion areas mainly includes metal surface corrosion detection based on traditional algorithms and metal surface corrosion detection based on deep learning. Metal surface corrosion detection based on traditional algorithms is mainly divided into threshold methods[13,14], morphological methods[15,16], edge detection method[17,18], and texture analysis[19–21]. These traditional methods are simple and easy to implement, but they generally have the disadvantages of low detection accuracy and being susceptible to background noise. In contrast, deep learning methods can automatically extract corrosion features from images, thus ensuring high-precision segmentation of the detected object. Liu et al.[22] used the VGG19 model to extract the damage and corrosion characteristics of steel plate coating, and combined it with the Faster R-CNN algorithm for defect detection. Srivastava et al.[23] used the UNet-8 layer architecture to analyze and classify corroded areas in images. Fondevik et al.[24] used PSPNet and Mask R-CNN for automatic image analysis of the self-built corrosion dataset. Tan et al.[25] built the DSNet model based on YOLOx, and demonstrated its accurate detection accuracy and segmentation performance in tunnel bolt detection and corrosion region segmentation tasks. Huang et al.[26] used the idea of ShuffleNetv2 to build a lightweight residual deep learning model, which achieved significant accuracy improvement in the metal surface corrosion segmentation task. Yin et al.[27] combined edge information and class balance loss with UNet to build a DeepSC-Edge model, and achieved excellent segmentation accuracy on the self-built corrosion image dataset. The above methods demonstrate the superior performance of the image semantic segmentation method based on CNN in metal corrosion detection. However, CNN has a limited receptive field and lacks contextual understanding when extracting features of pixels, resulting in low accuracy of image semantic segmentation, which affects the detection accuracy of metal surface corrosion.

Edge detection

In image segmentation, the edge information of the measured object plays a crucial role. The edge information can help the segmentation algorithm to capture the boundary and contour of the measured object more accurately, improving the precision and detail of segmentation. Xiao et al.[28] generated significant refined boundary information from the rough outline obtained by the Canny detector and integrate it into the BASeg framework to obtain fine boundary information. Zhu et al.[29] used the Sobel operator to design edge spatial attention blocks enhancing edge features and gradually extract edge features. Jagadeesh et al.[30] obtained edge features of varying granularity by subtracting average pooling values of different sizes from local convolutional feature maps and using these features at each step of the encoder to obtain reliable edge information. Tsai et al.[31] used the traditional Sobel edge detection algorithm to obtain a large amount of edge data to assist in model training, effectively enhancing feature representation without compromising inference speed. Bui et al.[32] constructed a multi-scale edge-guided attention network using Laplace operators to enhance the segmentation of weak boundary targets by preserving high-frequency edge information. Ma et al.[33] proposed the Laplacian operator-based active contour model (LOACM), which effectively enhances image edge detection capabilities

by utilizing the edge-sensitive properties of the Laplacian operator. Li et al.[34] constructed a dual extraction network model that uses the Laplacian operator to reduce edge blurring in underwater images, thereby enhancing edge clarity. Guo et al.[35] used four traditional edge detection algorithms—Canny, Sobel, Roberts, and Prewitt—to preprocess input images. The extracted edge features were used as input for an LVQ (Learning Vector Quantization) network with dynamic learning capabilities[36], making the predictions at the edges more accurate. Li et al.[37] used six gated mechanism modules to construct edge detection branches in the encoder to provide edge information for semantic segmentation and improve segmentation performance. Based on the above methods, this paper combines the traditional edge detection algorithm with multi-scale convolution to construct an edge-feature extraction module to improve the edge segmentation accuracy of the detected object.

Semantic segmentation method based on transformer

Inspired by the success of the Transformer in natural language processing, some researchers have applied this model and its variants to image semantic segmentation research[38–40]. Transformer models have the advantage of global perception, allowing them to consider pixels from all positions in an image simultaneously during semantic segmentation tasks. They effectively capture long-range dependencies between pixels, achieving a comprehensive and high-level understanding of the entire image. This, to a certain extent, compensates for the shortcomings of CNN models in handling long-range dependencies. Dosovitskiy et al.[41] applied the hierarchical Transformer structure ViT model to image classification for the first time. Zheng et al.[38] proposed the SETR model based on the ViT model, replacing CNN with pure Transformer as the encoder, regarded semantic segmentation as a sequence-to-sequence prediction task, and achieved 48.64% mIoU on the ADE20K dataset with single-scale inference. Strudel et al.[42] proposed Segmenter, which uses a ViT-based encoder and mask Transformer decoder to generate per-pixel class labels from attention maps between image patches and class embeddings. However, its backbone network can only output features at a single scale and has a high computational cost. To overcome this issue, The current solutions[43–45] mainly use hierarchical structures to extract features of different scales and introduce local computation to reduce the computational complexity of the model. However, these improvement methods mainly focus on the encoder design of the Transformer, while the role of the decoder has been largely overlooked.

Xie et al.[46] proposed SegFormer, which uses a hierarchical Transformer encoder without position coding to output fine-grained features with high resolution and coarse-grained features with low resolution at the same time, to better capture local and global information in images. A lightweight MLP (Multilayer Perceptron) decoder aggregates different levels of information to produce a simple, intuitive yet powerful representation. By layering the feature map and lightweight model, SegFormer achieves the characteristics of high precision, low computational cost, and small model volume, which gives it more prominent advantages and potential in metal surface corrosion detection tasks. However, due to the blurred boundary and irregular shape of the metal corrosion image, the SegFormer model has some problems such as the loss of details, texture, and edge segmentation errors during pixel segmentation. Therefore, it is necessary to further study and use edge features to segment metal corrosion images more accurately.

Methods

Network architecture

Given the problem that the SegFormer network frequently uses the Transformer self-attention mechanism, which causes the network to pay too much attention to high-level abstract information and ignore shallow low-level information such as edge and texture in the image, this paper uses the edge-feature extraction module to construct spatial branches within the network for extracting edge information from images. Since the shallow network contains rich edge information, the auxiliary branch built by the EEM module is used in this paper as the skip connection layer between the transformer block 1 module and the decoding layer in the SegFormer backbone. The auxiliary branch accepts the low-level semantic information from the output of the first layer of the encoder, extracts the edge features of the image, and provides these features to the decoder to restore the feature edge information of the image. The SegFormer metal corrosion image segmentation network based on parallel edge-feature extraction is shown in Fig. 1.

In the design, the EEM module is placed only in the first layer because the feature map at this layer contains the richest and most comprehensive edge information. Due to the multi-layer feature extraction mechanism of the Transformer, global semantic information becomes progressively stronger as the network depth increases, while local details, such as edge features, are gradually diluted or overwritten. Therefore, introducing edge information into the FFM of other layers may not be as effective as in the first layer. By directly passing the edge information extracted by the EEM to the FFM in the final layer, the model can fully utilize these fine-grained edge features during the decoding stage, thus significantly improving segmentation accuracy. Moreover, the final FFM layer receives edge features from the EEM, allowing the model to enhance its detail modeling capability based on global semantic information, without the need to frequently introduce edge features into intermediate layers. This design avoids interference from edge information in the consistency of global semantic modeling in intermediate layers, while significantly reducing computational costs and model complexity. By introducing edge features only in the final layer FFM, the design strikes the optimal balance between model performance and computational overhead.

Edge-feature extraction module (EEM)

The Sobel operator has the characteristics of simple calculation, fast speed, easy implementation, and effective detection of image edges, while the Laplacian operator can enhance the finer details within the image. Based on the advantages of the two operators, this paper uses the Sobel operator, multi-scale convolution operator, and Laplacian operator to construct a parallel edge-feature extraction module. Branch 1 of the edge-feature

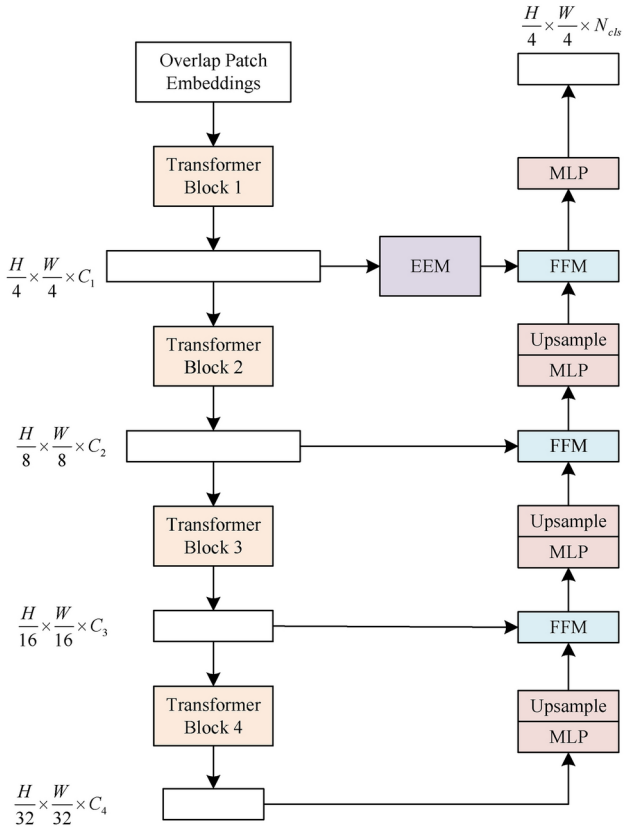

Fig. 1. SegFormer network optimization structure for parallel extraction of edge features.

extraction module highlights the edges of the detected region through Sobel edge detection and captures the features of the detected object at different scales by multi-scale convolution. Branch 2 uses the Laplacian operator to optimize the design of parallel structures and enhance the module's ability to extract details. By fully combining the advantages of the Sobel operator and the Laplacian operator, the network can capture the edge features of the detected object more comprehensively and extract more abundant details, and improve the segmentation precision and accuracy of the corroded image.

The edge-feature extraction module of this paper is shown in Fig. 2. The parallel structure of the Sobel operator and the Laplacian operator is used to construct the auxiliary branch of the SegFormer backbone network. The edge operator module is configured as a fixed kernel in the convolutional layers so that the model can automatically learn the edge information from the image. In the edge-feature extraction module, the input features of each branch are first adjusted by a 1×1 convolution layer, then an edge filter composed of two operators is used to extract edge features. For the Sobel edge detection branch, three depth separable convolution layers of different scales ($3 \times 3, 5 \times 5, 7 \times 7$) are joined in parallel. At the same time, the number of parameters and computational complexity are reduced while the features of different scales are captured. Finally, a 1×1 convolution layer is used for channel scaling. The process of the edge-feature extraction module is expressed as follows:

$$F_{Sobel} = f_{Conv1}(F_X) \otimes (G_{Sobel} * S_{Sobel}) + B_{Sobel}, \tag{1}$$

$$F_1 = f_{Conv1}([f_{DSConv3\times3}(F_{Sobel}), f_{DSConv5\times5}(F_{Sobel}), f_{DSConv7\times7}(F_{Sobel})]), \tag{2}$$

$$F_{Lp} = f_{Conv1}(F_X) \otimes (G_{Lp} * S_{Lp}) + B_{Lp}, \tag{3}$$

$$Y_E = f_{Conv1}([F_1, F_{Lp}]), \tag{4}$$

where $f_{\mathrm{Conv1}}(\cdot)$ represents 1×1 convolutional layer; $f_{\mathrm{Conv}\,3\times3}$, $f_{\mathrm{Conv}\,5\times5}$ and $f_{\mathrm{Conv}\,7\times7}$ represent 3×3, 5×5 and 7×7 depth separable convolution; F_X represents input signal; $\otimes$ represents the depth-wise convolution; G_{Sobel} and G_{Lp} represent Sobel filter and Laplacian filter; *denotes the channel-wise broadcasting multiplication; S_{Sobel} and S_{Lp} represent the scaling parameters; B_{Sobel} and B_{Lp} denote bias in the convolutional layer; F_{Sobel}, F_1 and F_{Lp} reprobel operator, multi-scale convolution and Laplacian operator respectively. Y_{E} refers to the edge features obtained after 1×1 convolution layer concatenation and fuses the features extracted by two branches.

In the metal corrosion segmentation task, the model needs to accurately segment the corrosion region from the background, which places high demands on edge detail modeling and the capture of multi-scale features.

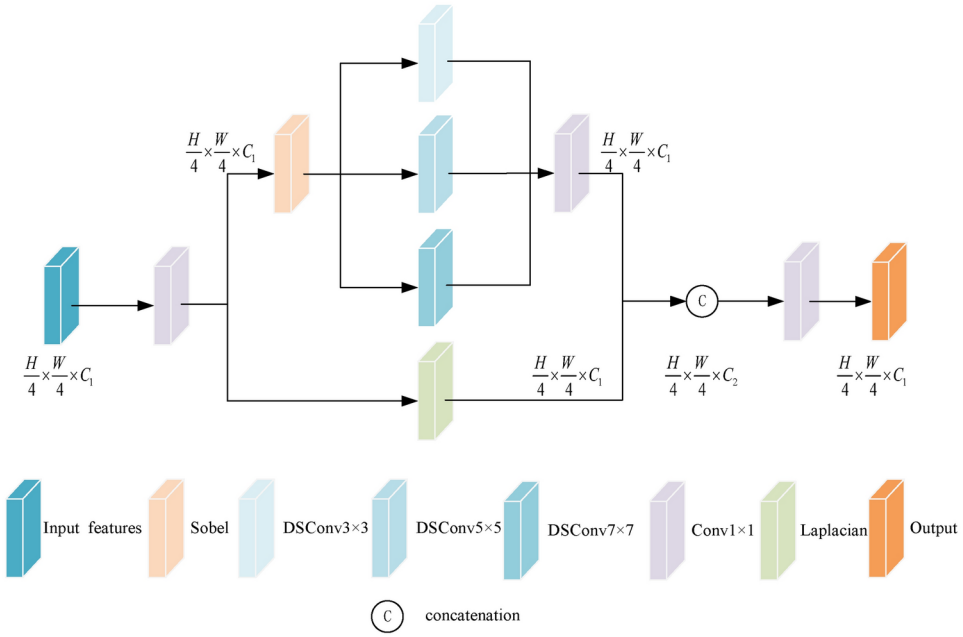

Fig. 2. Architecture of EEM.

To address this, the EEM module combines the advantages of the Sobel operator and the Laplacian operator to extract the edge features of the corrosion area from both local and global perspectives.

The Sobel operator captures local edge details by calculating the direction and magnitude of grayscale changes. Since the boundaries of metal corrosion regions exhibit significant scale differences, single-scale edge extraction is insufficient to describe the complex corrosion features. Therefore, multi-scale depthwise separable convolutions (3×3, 5×5 and 7×7) were introduced into the Sobel branch to effectively enhance the model's ability to perceive edge features at different scales. This also enables a comprehensive description of the corrosion region's edge details by merging multi-scale information. This design not only improves the model's robustness but also reduces computational overhead through depthwise separable convolutions. In contrast, the Laplacian operator, based on the second-order derivative property, is primarily used to capture global information about the overall edge structure. Compared to the Sobel operator, the Laplacian focuses more on the global shape characteristics of the boundary rather than on fine-grained local details. Due to its strong global modeling capability, there is no need to further model local features through multi-scale convolutions. This design simplifies the model structure while avoiding unnecessary computational overhead.

Finally, the multi-scale edge details extracted by the Sobel branch are fused with the global edge structure extracted by the Laplacian branch through the channel scaling mechanism. This edge feature fusion allows the model to complement one another in fine-grained boundary modeling and global shape understanding, thus improving segmentation performance. The Sobel branch accurately captures complex edge details and features at different scales, while the Laplacian branch provides supplementary global semantic boundary information. The synergistic effect of both branches significantly enhances the model's performance in metal corrosion segmentation tasks.

In SegFormer, the multi-layer structure of the Transformer often causes local details to be gradually overshadowed by global information during transmission, leading to the loss of details such as edges. The edge information in the first layer is typically more abundant because the local details in the feature map have not yet been excessively influenced by global information. By introducing an EEM module between Transformer Block 1 and the decoder, the model can effectively extract rich edge features. This not only enhances the representation of edge information but also ensures that this information is preserved and further refined at subsequent levels. Placing the EEM module at this position allows for the effective integration of global information extracted by the Transformer and fine-grained edge features during the decoding phase, aiding in the accurate capture of complex boundaries and small object details. This configuration enables the model to retain a global contextual understanding while precisely capturing local details in the image segmentation process. It is especially beneficial when handling complex scenes or images with blurred boundaries, as it significantly improves segmentation accuracy and performance.

Feature fusion module (FFM)

The output of the original SegFormer encoder contains low-level detailed features such as edges and textures, and high-level semantic features such as target shapes and classes. However, its decoder layer simply combines local attention with global attention, and aggregates multi-scale information from different coding layers, such as simple dimensional concatenation and upsampling, It cannot fully integrate high-level semantic information

and low-level detail information, and restore multi-scale features extracted by the coding layer, resulting in the loss of target feature information. In addition, the original SegFormer decoder mainly relied on MLP (Multilayer Perceptron) and simple convolution operations to fuse features from different levels. While this design is simple and efficient, it introduces additional parameters and computational overhead when fusing multi-scale features, especially when handling high-dimensional feature maps.

To effectively recover high-resolution output from the features transmitted by the encoder, this paper adopts a step-up sampling decoding layer design and introduces a feature fusion module (FFM) to achieve hierarchical and progressive feature fusion. At each level of the decoder, the feature map after the reduced resolution is fused with the high-resolution feature map of the corresponding level in turn, and upsampling is performed step by step to reduce the loss of detail information while fusing multi-scale feature information, to strengthen the feature propagation and fusion, so that the network can better adapt to the targets and details of different scales, and thus improve the segmentation performance of the network. To further reduce the number of parameters and computational complexity, the feature fusion module (FFM) adopts depthwise separable convolutions, decomposing standard convolutions into depthwise and pointwise convolutions. By employing simple addition and concatenation operations, the module effectively fuses feature maps of different scales. Through these efficient convolutional operations and feature fusion, the FFM ensures reasonable control of the spatial dimensions and number of channels in the feature maps, while preserving key information and reducing computational overhead.

The feature fusion module of this paper is shown in Fig. 3. First, low-level semantic features and high-level semantic features containing edge information are summed to obtain preliminary semantic features, and edge details and global context information are integrated to make up for the shortcomings of a single feature in semantic expression. The initial semantic features are spliced with the original low-level and high-level semantic features to expand the feature space, so that the model can capture more semantic information and form a rich feature representation. In order to improve efficiency, deep separable convolution is applied to the spliced features to extract features, which can effectively capture multi-scale information and significantly reduce the computational cost. The 1×1 convolution layer and sigmoid activation function are then used to generate an attention map, which is multiplied by elements with the original low-level and high-level semantic features to further enhance the edge feature representation. This attention mechanism allows the model to focus on key feature areas and filter out irrelevant background and noise. Finally, the processed features are spliced and fused, and the final fusion features are obtained through 1×1 convolution. This process effectively combines edge information, global context information and multi-scale features to ensure the optimal balance between semantic integrity and detailed description of the output features.

$$F' = F_{Edge} \oplus F_C, \tag{5}$$

$$F_A = \sigma\left(f_{Conv1}\left(f_{DSConv3\times3}\left[F_{Edge}, F'\right]\right)\right), \tag{6}$$

$$F_B = \sigma\left(f_{Conv1}\left(f_{DSConv3\times3}\left[F_C, F'\right]\right)\right), \tag{7}$$

$$Y = f_{Conv1}\left(\left[F_A \odot F_{Edge}, F_B \odot F_C\right]\right), \tag{8}$$

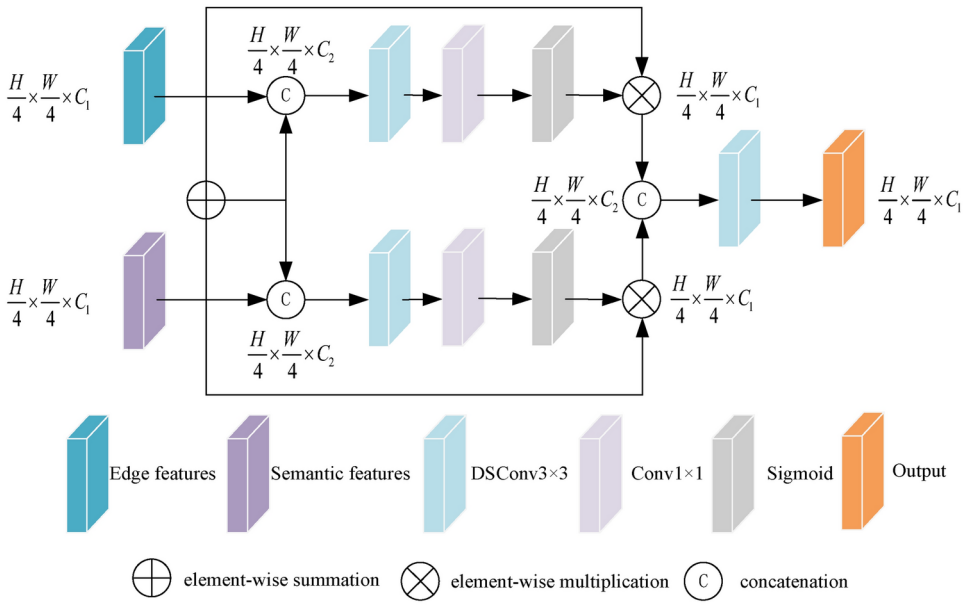

Fig. 3. The architecture of the feature fusion module (FFM).

where F_{Edge}, F_{C} and $F^{'}$ represent low-level semantic features containing edge information, high-level semantic features and aggregated preliminary semantic features respectively, $f_{\mathrm{Conv1}}(\cdot)$ denotes a 1×1 convolution layer; $f_{\mathrm{DSConv3\times3}}$ represents 3×3 depth separable convolution; σ denotes sigmoid activation function; $\odot$ represents element-by-element multiplication; F_{A} and F_{B} denotes the generated attention graph; Y represents the final fusion feature obtained by concatenated and fused the processed features and then passing through the 1×1 convolution layer.

Compared to the original SegFormer decoder, the FFM module, by introducing and enhancing edge information, is better suited for modeling complex boundaries and fine-grained features. By combining low-level and high-level semantic features, it achieves collaborative modeling of global semantic relationships and local edge details, significantly improving segmentation accuracy. Meanwhile, the application of depth separable convolution significantly reduces computational costs while ensuring feature extraction capabilities, thereby enhancing the model's efficiency. Additionally, the effective extraction of multi-scale information makes the module more robust to target regions with significant scale variations, such as corrosion pits, allowing for more accurate segmentation.

Experiments

This section introduces the dataset, evaluation metrics, experimental setup, loss function, and hyperparameter settings. Then, comparative experiments are conducted with different models, and the experimental results are analyzed. Finally, tests are performed to verify the robustness and generalization capabilities of the proposed model.

Dataset

Three datasets were used in this experiment: the metal surface corrosion image dataset produced by the University Libraries Virginia Tech[47], the self-built tubing corrosion pit image dataset, and the Industrial Machine Tool Element Surface Defect Dataset[48].

Metal surface corrosion public image dataset

The metal surface corrosion image dataset produced by the University Libraries Virginia Tech is semantically annotated based on the Bridge Inspector's Reference Manual (BIRM) and the Corrosion Status guidelines of the American Association of State Highway and Transportation Officials (AASHTO). The dataset contains 440 images and is classified into four categories: Background, Fair, Poor, and Severe.

Aiming at the lack of original data, the insufficient amount of training data negatively impacts the performance of image semantic segmentation. In this paper, the dataset is divided into the training set and the test set according to the ratio of 9:1. Each image was adjusted to a 512×512 pixel small image, allowing a 0.5 overlap ratio between the small images. Given the problem that the local information of the incorrectly labeled image is amplified due to slice processing, the abnormal data is removed when used. Finally, themetal surface corrosion image dataset is obtained, which consists of 5121 training sets, 570 validation sets, and 338 test sets.

Self-built tubing corrosion pit image dataset

The self-built tubing corrosion pit image dataset is collected from the oil field pipelines and photographed under different lighting, angles, and other conditions to ensure the diversity of the data. After manual filtering, a total of 227 images were retained. Data annotation is done by three professionally trained graduate students using Labelme annotation tools to ensure accuracy and consistency. The completed tubing corrosion pit image and its corresponding label are shown in Fig. 4, where the red mark represents the area of the corrosion pit. After the same image processing method, 1230 training images, 137 validation images, and 128 test images are obtained.

The industrial machine tool element surface defect dataset (BSData)

The BSData dataset contains 1104 images with 3 channels and 394 annotated images for the surface damage type 'pitting'. In this study, 394 annotated images were used. After applying the same image processing methods, the final pitting image dataset consists of 621 training images, 70 validation images, and 72 test images, with a size of 448×448 pixels for the processed images.To describe the class distribution of the three datasets used, Table 1 shows the corrosion ratios for each dataset, where the corrosion ratio refers to the ratio of corrosion pixel count to the total pixel count in an image.

Evaluation metrics

In this paper, mean Intersection over Union (mIoU), Accuracy (Acc), model parameters (Params), Floating Point Operations per Second (Flops), and Frames Per Second (FPS) were evaluated. The average intersection over union (IoU) is the mean value of the IoU across all categories, used to evaluate the accuracy of the model segmentation results. Accuracy is calculated by dividing the number of correctly predicted samples by the total number of samples, used to measure the model classification performance. The number of model parameters is calculated by counting all the learnable parameters within the model, used to assess the computational complexity of the model. Flops refer to the number of floating-point operations the model can perform per second, used to measure the computational performance of the model. Frames per second (FPS) indicate the number of image frames the model can process and display in one second, used to measure the speed of the model segmentation. The model parameters are measured in MB, and the FLOPs are measured in GFLOPs (billion floating-point operations).

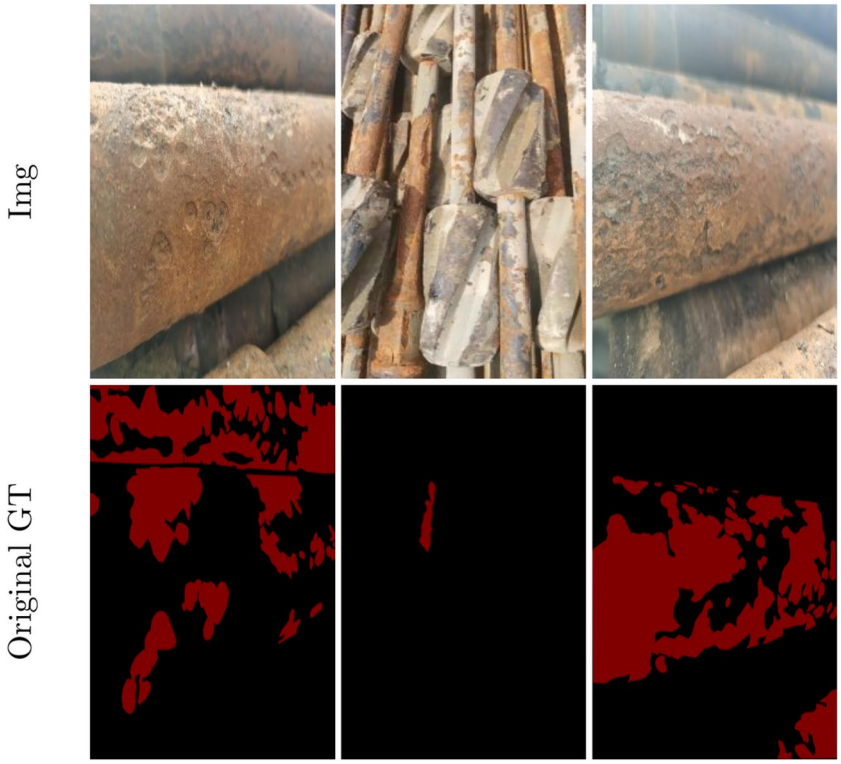

Fig. 4. Samples of our self-built tubing corrosion pit image dataset.

Dataset	Training and validation data				Test data			
	Corrosion ratio			Background (%)	Corrosion ratio			Background (%)
	Fair	Poor	Severe		Fair	Poor	Severe	
(1)	21.00%	10.90%	2.40%	64.50	19.90%	12.80%	1.70%	67.20
(2)	17.18%			82.82	18.90%			81.10
(3)	1.27%			98.73	1.34%			98.66

Table 1. Class distribution of the dataset used in the current work. (1) Metal surface corrosion public image dataset; (2) Self-built tubing corrosion pit image dataset; (3) The Industrial Machine Tool Element Surface Defect Dataset

Loss function

Focal loss is used to address the class imbalance in the metal corrosion dataset. Due to the imbalance between positive and negative samples, misclassification is common during training. Focal loss effectively alleviates the problem of difficult sample training by adjusting the weights of the loss function.The loss is defined as follows:

$$L_{\text{Focal}}(p_t) = (1 - p_t)^{\gamma} \log(p_t). \tag{9}$$

Focal loss reduces the loss for easily classified samples by introducing a tunable scaling parameter γ, which focuses the model's training on difficult samples. This allows the model to concentrate more on learning the features of hard-to-classify samples, improving classification accuracy on these challenging instances. In this paper, γ is set to 2. p_t represents the ratio of the number of pixels of the target class to the total number of pixels in the image.

Impact of hyperparameters

This section discusses the impact of hyperparameters, with results shown in Table 2. Batch size and learning rate are critical hyperparameters in deep learning training. By properly adjusting these two hyperparameters, the training efficiency and performance of the model can be effectively improved.

All experiments were performed on the Linux 5.4.0-126-generic 64-bit operating system. The hardware configuration includes an Intel(R) Xeon(R) Platinum 8255C CPU with 12 vCPUs and two NVIDIA GPUs: an RTX 3090 (24GB) GPU for training and an RTX 3060 GPU for inference. The experiments were conducted using the PyTorch deep learning framework, with the development environment consisting of Python 3.8, PyTorch

Learning rate	Batchsize	Epoch	Training loss	Validation loss	mIoU (%)
0.001	8	300	0.047	0.061	91.15
0.001	16	300	0.042	0.056	91.2
0.0001	8	300	0.048	0.06	91.16
0.0001	16	300	0.038	0.054	91.41

Table 2. Result of impact discussion of hyperparameters.

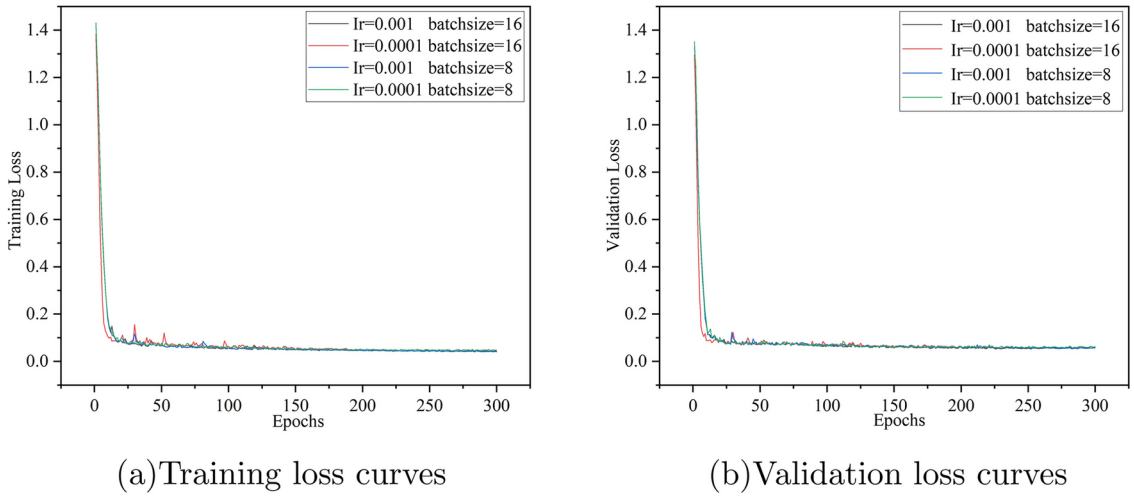

Fig. 5. Class distribution of the dataset used in the current work.

1.13.1, and CUDA 11.6. The proposed model was trained on the following datasets: the metal surface corrosion image dataset, the self-built tubing corrosion pit image dataset, and the BSData dataset. When using the RTX 3090 for training, the model training times were approximately 26 h for the metal surface corrosion image dataset, 8 h for the self-built tubing corrosion pit image dataset, and 4 h for the BSData dataset. During inference with the RTX 3060, the inference time per image was approximately 14–16 ms. Adaptive Moment Estimation (Adam) was used for model training. The cosine annealing strategy is employed to adjust the learning rate, and a momentum of 0.9 is used to update the network parameters. The training consists of 300 epochs, with a batch size set to 16 and an initial learning rate of 0.0001. The results of the hyperparameter impact analysis are shown in Table 2. Training plots (showing both training loss and validation loss) have been generated. In Fig. 5, a smaller learning rate of 0.0001 allows for more fine-grained optimization of the model, demonstrating a more stable training process and lower validation loss. On the other hand, the larger batch size of 16 yields better performance across training loss, validation loss, and mIoU, indicating that it facilitates more stable gradient updates, thus improving overall model performance. Moreover, the experiment used 300 training epochs, which is sufficient to ensure that the model converges on the dataset, and the training process remains stable. The model achieves the best performance when the learning rate is 0.0001, the batch size is 16, and the number of training epochs is 300.

Comparison method selection

The comparison models used in this paper include CNN-based models such as U-Net, HRNet, and DeepLabv3+, as well as Transformer-based models such as SETR, Segmenter, Swin Transformer, and SegFormer. CNN-based models excel in handling local details and fine-grained features in metal corrosion segmentation tasks. Among them, U-Net adopts a U-shaped network structure with downsampling and upsampling paths. The encoder part uses convolutional and pooling layers to extract high-level semantic features by gradually reducing the size and channels of feature maps. The decoder part employs transposed convolutional layers and skip connections to restore the size and channels of feature maps while fusing low-level features from the encoder with high-level features from the decoder to recover spatial details and achieve precise segmentation. In metal surface corrosion segmentation tasks, U-Net extracts contextual information through the encoder and performs accurate localization through the decoder. HRNet adopts a parallel multi-resolution subnetwork design and repeated multi-scale fusion operations, enabling each high-to-low resolution representation to repeatedly receive information from other parallel representations, thereby obtaining rich high-resolution representations. Through parallel multi-resolution convolutions and repeated multi-resolution fusion, HRNet maintains high-resolution features and effectively integrates cross-scale information, making it suitable for fine-grained segmentation tasks such as metal surface corrosion segmentation. DeepLabv3+ consists of an encoder and a decoder. The encoder uses a backbone network with dilated convolutions to extract features from the input image and integrates multi-scale

contextual information through the ASPP module, enhancing the model's ability to segment objects at different scales. The decoder gradually restores spatial resolution through upsampling and fuses low-level and high-level feature maps to produce high-quality semantic segmentation results. It primarily relies on the pyramid pooling module, which extracts image features through convolutions at different dilation rates and multiple effective receptive fields, and encodes multi-scale contextual information through pooling operations, thereby improving the model's accuracy in metal surface corrosion segmentation tasks. Unlike traditional convolutional neural networks, SETR abandons convolutional operations and employs a pure Transformer encoder to extract global semantic information from images. By treating semantic segmentation as a sequence-to-sequence prediction task, SETR better captures global contextual relationships in images. Its Transformer encoder models global semantic information and captures long-range dependencies, enabling it to better handle the complex textures and distributions of corrosion regions while demonstrating strong robustness and scalability. The Segmenter model shares some similarities with SETR in its basic framework, as both treat semantic segmentation as a sequence-to-sequence problem and use Transformer structures for image segmentation. While SETR focuses on feature extraction using Transformers, Segmenter innovatively incorporates object queries and category embedding mechanisms, leveraging cross-attention operations to more finely process image patches and category information, generating pixel-level category labels. Its self-attention mechanism allows the model to directly model interactions between any pixels in the image, overcoming the limitations of traditional convolutional neural networks in local receptive fields, thereby more accurately identifying the complex textures and distributions of corrosion regions. Swin Transformer divides the image into multiple patches and performs self-attention calculations within each window, avoiding the high computational cost of global self-attention. Additionally, the model adopts a hierarchical structure similar to convolutional neural networks, gradually extracting features at different scales through layer-by-layer downsampling. This design not only improves computational efficiency but also preserves the Transformer's ability to model long-range dependencies. By introducing sliding windows and hierarchical self-attention mechanisms, Swin Transformer efficiently processes high-resolution images in metal surface corrosion segmentation tasks, accurately capturing the complex textures and distributions of corrosion regions, while its hierarchical structure enables the model to balance local details and global information.

The images of metal surface corrosion exhibit significant common features, such as irregular textures, color variations, changes in brightness and local contrast, and uneven distribution of corrosion areas. SegFormer, a lightweight semantic segmentation model, can effectively capture these subtle changes through adaptive feature extraction and cross-layer feature fusion, enabling accurate localization and identification of the corrosion areas. Its hierarchical Transformer encoder combined with a simple and efficient All-MLP decoder allows SegFormer to extract multi-scale features precisely while reducing computational overhead, making it particularly suitable for small and complex corrosion regions. Meanwhile, corrosion areas are often affected by uneven lighting and image noise. SegFormer can better model multi-scale features and improve the model's robustness in complex environments by incorporating multi-scale feature fusion and global contextual information. Moreover, SegFormer ensures high segmentation accuracy in high-resolution image processing while maintaining efficient computation and real-time inference performance, giving it a strong application advantage in metal corrosion detection tasks. It can accurately and robustly handle various types of corrosion images.

Models based on CNNs, such as U-Net, HRNet, and DeepLabv3+, can handle local details and fine features well in the metal corrosion segmentation task. However, they have certain limitations in modeling long-range dependencies. Transformer-based models, like SETR, Segmenter, and Swin Transformer, excel at capturing global context and long-range dependencies but often overlook small local features, and they tend to have high computational complexity, resulting in slower inference speeds, which makes them less suitable for real-time applications. In contrast, SegFormer combines the strengths of both CNNs and Transformers, allowing for precise segmentation of corrosion areas while maintaining efficient computation and real-time inference capabilities. Therefore, this paper aims to improve SegFormer and compare it with traditional CNN-based models (such as U-Net, HRNet, and DeepLabv3+) and Transformer-based models (such as SETR, Segmenter, and Swin Transformer). Through experiments, the paper demonstrates the performance advantages of the improved SegFormer in the metal surface corrosion segmentation task, especially in terms of global context modeling, local detail capture, computational efficiency, and inference speed.

Experiment

To validate the segmentation accuracy of the SegFormer model, which is based on parallel edge-feature extraction in the metal surface corrosion image detection task, this paper utilizes the edge-feature extraction module as an auxiliary branch of the backbone network and places it between the Transformer Block 1 module and the decoding layer as a skip connection layer aimed at capturing and transmitting edge information within the images. In addition to introducing the auxiliary branch, the decoding layer has been optimized with the gradual upsampling design. The feature fusion module is introduced to achieve hierarchical feature fusion by combining features of different scales, capturing local detail changes while preserving the global context of the image. To validate the effectiveness of the proposed model, this paper utilizes the Metal Surface Corrosion public image dataset, the self-built tubing corrosion pit image dataset, and the BSData dataset for experiments. The proposed algorithm is compared and validated against other algorithms using the evaluation metrics detailed in the previous section. The evaluation metrics, experimental environment, and parameter settings mentioned in this chapter have already been elaborated in the previous chapter, and thus will not be repeated here.

Model	mIoU ↑	Acc ↑	FPS ↑	Params ↓	Flops ↓
SegFormer	66.20	85.32	**66.37**	**3.72**	**6.78**
SegFormer-Sobel	66.07	85.96	52.96	3.76	9.55
SegFormer-Laplacian	66.11	86.13	55.95	3.78	9.09
SegFormer-EEM	**67.45**	**86.53**	48.12	3.87	9.13

Table 3. Ablation experiment of edge operator in EEM. Significant values are in bold.

Model	mIoU ↑	Acc ↑	FPS ↑	Params ↓	Flops ↓
SegFormer	66.20	85.32	66.37	3.72	6.78
SegFormer-LawinASPP[49]	59.33	81.36	34.95	**3.45**	3.09
SegFormer-FFM	**66.25**	**86.03**	**69.81**	3.58	**2.79**

Table 4. Ablation experiment with decoder. Significant values are in bold.

Experiments based on metal surface corrosion public image dataset
Firstly, ablation experiments were conducted on the metal surface corrosion public image dataset to analyze the effectiveness of each component of the proposed method, and then the performance of the proposed method was compared to other methods to verify the superiority of the proposed method.

Impact of edge operator In this paper, SegFormer MiT-B0 is used as the baseline model to analyze the validity of the three-branch edge detection operator in the edge-feature extraction module. According to the experimental results in Table 3, SegFormer introduced with the EEM module was 67.45% mIoU. Compared to the SegFormer models with single Sobel or Laplacian operator modules, the SegFormer-EEM shows improvements in mIoU by 1.38% and 1.34%, respectively. Compared to the baseline SegFormer, the mIoU is improved by 1.25% and accuracy (Acc) by 1.21%.

The experimental results show that the edge detection performance of individual Sobel and Laplacian operators is limited, and introducing a single edge operator as an auxiliary branch in the baseline SegFormer does not improve performance. In the EEM module, the Sobel operator extracts edge information, while the Laplacian operator extracts high-frequency and texture information from the image. The combination of these two operators provides richer and more diverse feature representations, thereby improving the detection performance of the corroded area.

Impact of decoder This paper employs different decoders to validate the effectiveness of the feature fusion decoding layer in the task of metal surface corrosion segmentation. In Table 4, SegFormer-Lawin refers to the decoder using the Lawin Transformer. The experimental results show that the SegFormer with the feature fusion decoding layer achieves 66.25% mIoU. Compared to the original MLP decoder and the Lawin Transformer decoder, the mIoU is improved by 0.05% and 6.92%, respectively, and accuracy (Acc) of 86.03% shows an increase of 0.71% and 4.67% respectively.

The experimental results show that the use of the feature fusion decoding layer has a certain improvement in each evaluation metric. However, the Lawin Transformer decoder adopts a large window attention mechanism, which cannot effectively capture the details, edges, and textures required for the metal surface corrosion segmentation task, resulting in poor performance. In contrast, SegFormer uses the feature fusion decoding layer to gradually recover the details of the input image by gradually upsampling, thus improving the accuracy of segmentation. In addition, it is worth noting that the number of parameters of the model in this paper is lower than that of the baseline SegFormer, and FFM can reduce the complexity and number of parameters of the model, making the overall model more lightweight.

Ablation experiments To verify the effectiveness of the edge-feature extraction module and feature fusion upsampling decoding layer in the SegFormer model based on parallel edge-feature extraction on metal surface corrosion segmentation tasks, MiT-B0 was selected as the backbone of SegFormer in this paper to conduct ablation experiments on the metal surface corrosion segmentation dataset. According to the experimental results in Table 5, this model is 67.69% mIoU, which is 1.49% higher than the baseline SegFormer, and the Acc is 86.56%, which is 1.24% higher than the baseline SegFormer. The improvement of the accuracy evaluation metrics indicates that the model in this paper can effectively improve the accuracy of metal surface corrosion image segmentation. In addition, it is worth noting that the number of parameters of the model in this paper is lower than that of the baseline SegFormer, and FFM can reduce the complexity and number of parameters of the model, making the overall model more lightweight.

Comparative experiments In Table 6, this paper compares the segmentation performance of other semantic segmentation models based on CNN and the Segformer model on the metal surface corrosion dataset, and the results show that the SegFormer model has significant advantages in several evaluation metrics. To further explore the performance of the proposed model, this paper replicates the segmentation results of four semantic

Baseline	EEM	FFM	mIoU ↑	Acc ↑	FPS ↑	Params ↓	Flops ↓
✓			66.20	85.32	66.37	3.72	6.78
✓	✓		67.45	86.53	48.12	3.87	9.13
✓		✓	66.25	86.03	**69.81**	**3.58**	**2.79**
✓	✓	✓	**67.69**	**86.56**	62.56	3.60	3.06

Table 5. Ablation experiment. Significant values are in bold.

Model	Backbone	mIoU ↑	Acc ↑	FPS ↑	Params ↓	Flops ↓
U-Net[50]	VGG-16	45.58	78.77	16.13	24.89	225.87
FCN[51]	ResNet50	53.89	82.43	22.75	35.31	148.54
HRNet[52]	HRNetV2-W18	56.16	83.96	21.75	9.64	18.67
DeepLabv3+[53]	MobileNetV2	57.69	83.05	61.22	5.81	26.44
UPerNet[54]	Swin-T	60.18	84.25	36.51	58.94	236.00
SETR PUP[38]	ViT-T	64.44	86.44	4.47	315.39	416.00
Segmenter[42]	ViT-S	65.06	86.27	35.70	25.98	37.37
Segformer[46]	MiT-B0	66.20	85.32	**66.37**	3.72	6.78
Ours	MiT-B0	**67.69**	**86.56**	62.56	**3.60**	**3.06**

Table 6. Evaluation results of different semantic segmentation models on metal surface corrosion public image dataset. Significant values are in bold.

segmentation models based on the Transformer on the metal surface corrosion dataset and compares them with the proposed model. Table 6 shows that the proposed model is 67.69% mIoU, which is 1.49% higher than that of the baseline SegFormer model and 7.51%, 3.25%, and 2.63% higher than that of UPerNet, SETR, and Segmenter, respectively. FCN and U-Net lose more detail and local information during downsampling and upsampling. Downsampling results in a reduction in image size and spatial resolution, thus ignoring some subtle corrosion features. In the process of upsampling, the lack of detailed information makes the edge and small structure unable to recover, resulting in a decline in the accuracy of the image segmentation results. The feature extraction ability of HRNet model is insufficient, and it is difficult to capture rich details and texture features in the corrosion image, so the segmentation effect is poor. DeepLabv3+ uses hollow convolution to expand the receptive field. However, due to the discontinuity of the receptive field of the hollow convolution, the details and texture features of the corroded region cannot be accurately captured, which limits the improvement of the model accuracy. The UPerNet, SETR, and Segmenter had better segmentation performance than the CNN model, but their number of parameters was tens to hundreds of times higher than that of the proposed model. This increase in the number of parameters leads to a significant increase in computational complexity, as well as a significant increase in memory and computational power requirements for training and reasoning. Nevertheless, the proposed model strikes a balance between performance and efficiency, and can maintain high segmentation accuracy while while maintaining a lightweight architecture, which is suitable for practical application scenarios with limited computing resources. This high efficiency makes the proposed method more practical in resource-constrained environments.

The inference speed of the above models was evaluated on a single NVIDIA GeForce RTX 3060 GPU, and the segmentation accuracy and speed of each model on the metal surface corrosion dataset are presented in Table 6. The model in this paper verifies the effectiveness of the proposed method with a competitive advantage of 3.60 MB and 3.06 GFLOPs.

The experimental results of the proposed model and other semantic segmentation models based on Transformer architecture on the metal surface corrosion dataset are visualized, as shown in Fig. 6. Figure 6 shows the excellent performance of the proposed method in metal surface corrosion boundary segmentation, which can accurately segment the corrosion boundary and correctly predict the types of metal surface corrosion. It can also be seen from the figure that the proposed method can effectively identify and segment small-area corrosion, while successfully suppressing the misclassification of fair corrosion as background. Compared with comparison methods, the proposed method not only has higher overall segmentation accuracy for metal surface image segmentation, but also performs better in the recognition and segmentation of small-area corrosion, and also has clearer segmentation boundaries for different corrosion categories.

Compared with other semantic segmentation models in this paper, the classification of metal surface corrosion and the segmentation of corrosion boundaries are poor, on the one hand, because the corrosion image categories in the public dataset used in the experiment are not balanced. In the dataset, the proportion of fair corrosion and poor corrosion is more, and the proportion of severe corrosion is less, which leads to the model tending to predict more fair corrosion and poor corrosion during the training process, so all models perform poorly in the segmentation of severely corroded regions. On the other hand, due to the complex features, irregular distribution, and unclear boundary of the metal surface corrosion image, the results of corrosion boundary segmentation are not accurate.

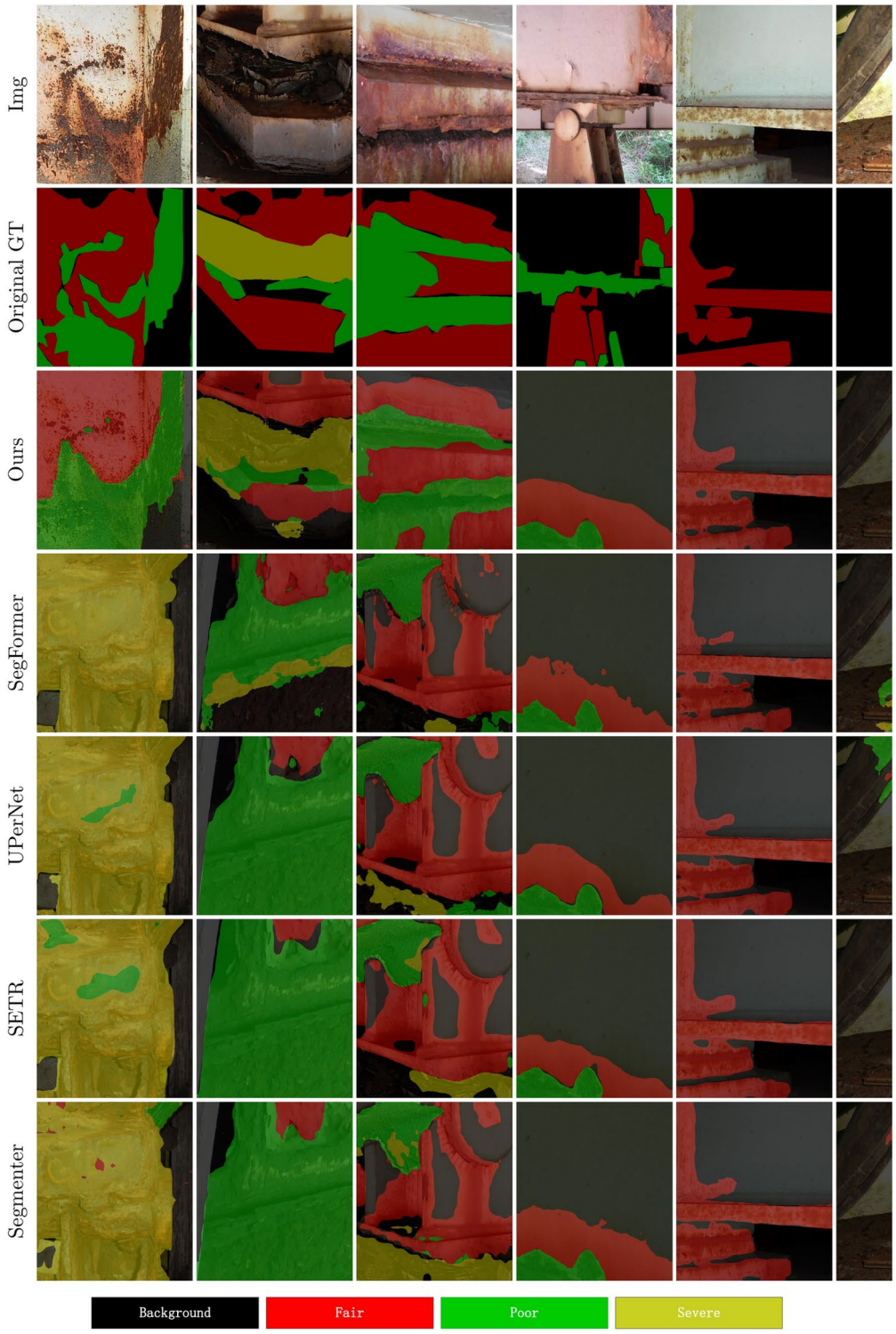

Fig. 6. Visualization results of different semantic segmentation models on metal surface corrosion public image dataset.

Experiments based on self-built tubing corrosion pit image dataset

To further verify the effectiveness of the proposed method, ablation experiments were performed on the self-built tubing corrosion pit image dataset, and the performance was compared with other methods to evaluate the applicability and advantages of the proposed method in different datasets and application scenarios.

Ablation experiments The ablation experiments of the proposed model on the self-built tubing corrosion pit image dataset in Table 7 further verify the effectiveness of the proposed edge-feature extraction module and fea-

Baseline	EEM	FFM	mIoU ↑	Acc ↑	FPS ↑	Params ↓	Flops ↓
✓			86.31	95.36	**62.06**	3.72	6.85
✓	✓		88.66	96.29	47.32	3.87	9.13
✓	✓	✓	**89.32**	**96.52**	61.74	**3.60**	**3.06**

Table 7. Ablation experiments of the proposed model on self-built tubing corrosion pit image dataset. Significant values are in bold.

Baseline	FFM	EEM layers	mIoU ↑	Acc ↑	FPS ↑	Params ↓	Flops ↓
✓			86.31	95.36	**62.06**	3.72	6.85
✓	✓	1	89.32	96.52	61.74	**3.60**	**3.06**
✓	✓	2	89.46	96.53	58.59	3.66	3.29
✓	✓	3	89.59	96.60	54.52	3.97	3.61
✓	✓	4	**89.67**	**96.65**	52.13	4.73	3.80

Table 8. Impact of branch number. Significant values are in bold.

Model	Backbone	mIoU↑	Acc↑	FPS↑	Params↓	Flops↓
U-Net[50]	VGG-16	83.41	94.4	16.24	24.89	225.87
HRNet[52]	HRNetV2-W18	84.81	94.86	27.68	9.64	18.67
DeepLabv3+[53]	MobileNetV2	85.39	95.16	60.07	5.81	26.44
UPerNet[54]	Swin-T	88.07	96.12	36.71	58.94	236
SETR PUP[38]	ViT-T	87.84	96	5.28	308	416
Segmenter[42]	ViT-S	85.81	92.86	34.85	25.97	37.37
TransUnet[40]	ViT-B	86.73	93.46	30.86	93.23	32.24
SwinUnet[55]	Swin-T	87.89	96.6	24.69	41.34	34.79
Segformer[46]	MiT-B0	86.31	95.36	**62.06**	3.72	6.85
Ours	MiT-B0	**89.32**	**96.52**	61.74	**3.6**	**3.06**

Table 9. Evaluation results of different semantic segmentation models on the self-built tubing corrosion pit image dataset. Significant values are in bold.

ture fusion upsampling decoding layer. According to the experimental results in Table 7, on the self-built tubing corrosion pit image dataset, this model is 89.32% mIoU, which is 2.99% higher than the baseline SegFormer, and the Acc is 96.52%, which is 1.16% higher than the baseline SegFormer.

Impact of branch number To verify the influence of the placement and layer number of the edge-feature extraction module on the model performance and computational complexity, this paper conducts an ablation experiment by adding multiple EEM layers between the encoder and decoder. The layer number of EEM represents adding the edge-feature extraction module in the transformer block 1, 2, 3, and 4 and the corresponding decoding layer. The experimental results, as shown in Table 8, demonstrate that the optimal balance between performance and complexity is achieved by adding a edge-feature extraction module between the transformer block 1 and the corresponding decoding layer, where the model maintains a lower complexity and achieves an mIoU accuracy of 89.32%. The edge feature information is mainly concentrated in the shallow layers of the main trunk transformer, and the edge feature information provided by subsequent layers decreases gradually. Therefore, the additional modules have limited impact on the overall performance improvement.

Comparative experiments The evaluation results of different semantic segmentation models on the self-built tubing corrosion pit image dataset are shown in Table 9, which is consistent with the experimental results obtained on the metal surface corrosion image dataset. Compared with the other semantic segmentation models based on CNN and Transformer, the proposed model achieves 89.32%mIoU accuracy with 3.60 MB of parameters.

The experimental results of the proposed method and the mainstream semantic segmentation model on the self-built tubing corrosion pit image dataset are visually compared, and the results are shown in Fig. 7. Figure 7 shows the advantages of the proposed algorithm in identifying and segmenting small-area corrosion pits, which can effectively detect the existence of corrosion pits and perform accurate segmentation. Especially when dealing with the boundary of corrosion pit, the proposed method can achieve high precision segmentation, which is obviously superior to other mainstream methods. Through comparison, it can be seen that the method presented

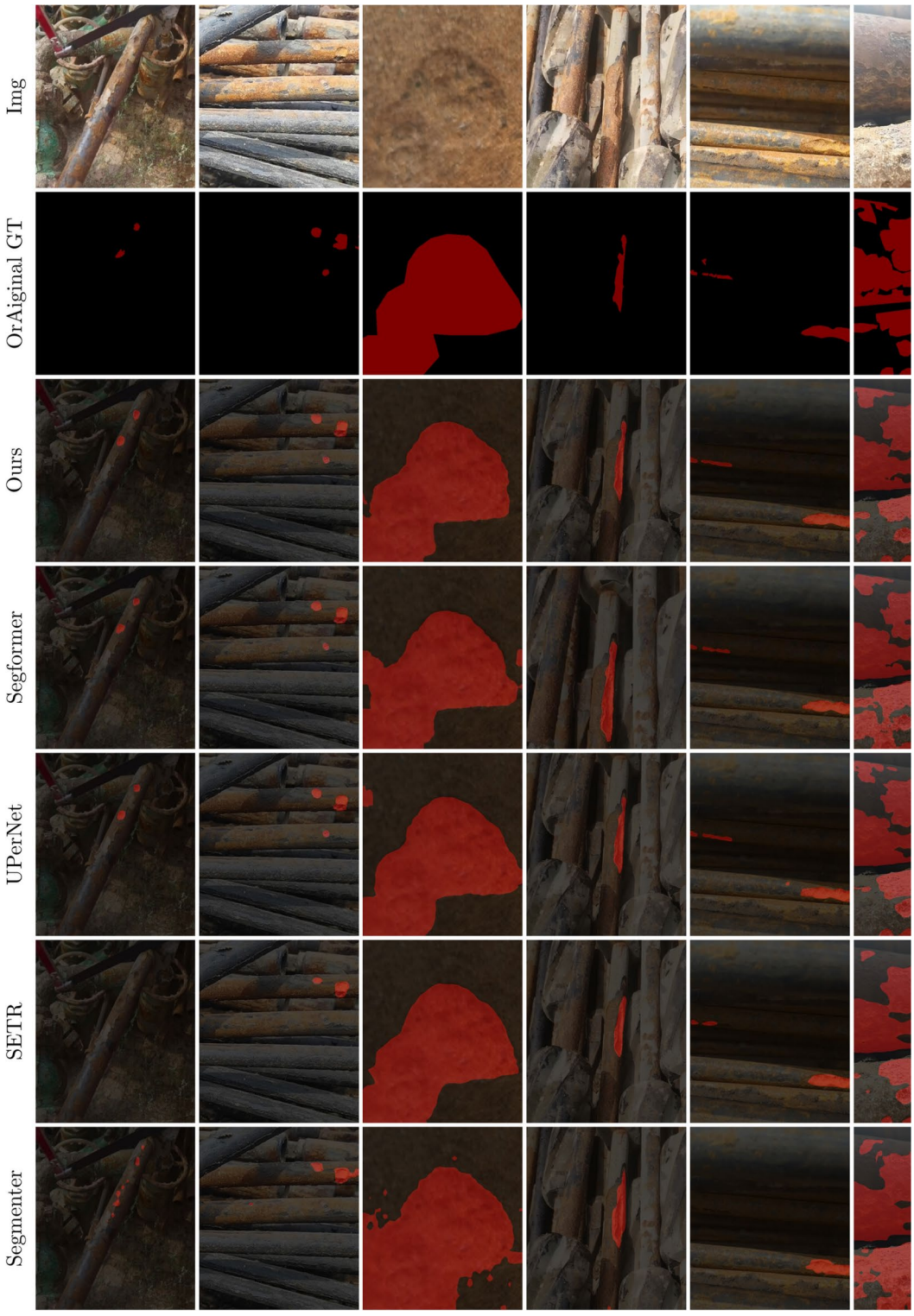

Fig. 7. Visualization results of different semantic segmentation models on self-built tubing corrosion pit images.

in this paper has strong capabilities in detail processing and accurate identification of small-range defects, and can better adapt to the complexity and diversity of tubing corrosion pit images.

Experiments based on BSData

Due to the limited number of samples of the self-built tubing corrosion pit image dataset, the reliability of model training is affected. This section introduces the publicly available BSData dataset, which contains surface damage

types similar to the "pitting" defect type in our dataset. By adding this dataset, the sample data can be enriched, and the credibility and reliability of the experimental results can be further enhanced.

Impact of multi-scale convolution in EEM To verify the effect of multi-scale convolution operations in the EEM, multi-scale convolution was added to the Sobel and Laplacian branches of the EEM respectively. It can be seen from the experimental results in the Table 10 that the model achieves the best effect when the Sobel branch is added with multi-scale convolution.

The experimental results further verified that the introduction of multi-scale depth separable convolution in the Sobel branch effectively enhanced the model's perception of edge features of different scales, while Laplacian operator paid more attention to the global shape features of the boundary. The combination of the two operators provided a richer and diversified feature representation, thus improving the detection performance of the corrosion region.

Comparative experiments To further validate the effectiveness of the proposed model,we conducted a comparative experiment on a publicly available BSData image dataset to evaluate the segmentation performance of the model in a small area of corrosion pits. As shown in Fig. 8, the SegFormer and SETR models performed poorly in predicting fine pitting regions, failing to accurately identify and segment these areas. UPerNet improved edge segmentation to some extent, but the overall segmentation performance was still unsatisfactory. Segmenter exhibited some smoothing effect on the edges, but there were still issues with insufficient segmentation detail. In contrast, the proposed method can accurately predict the edge of the corroded region, successfully segment the defect region, and ensure the high accuracy and integrity of the segmentation results (Table 11).

In addition to the comparison, the performance of the model in different indicators on the DSBata dataset was also counted. U-Net was based on the VGG-16 backbone network, with relatively low mIoU, a relatively large model and high computational complexity, resulting in slow inference speed. The multi-branch design of HRNet ensures high-quality local feature extraction, combined with powerful high-resolution information, which helps capture fine details. mIoU is slightly higher than U-Net, but its inference speed is slower. DeepLabv3+, combined with dilated convolution and spatial pyramid pool (ASPP), enhances the sensitive field and can better capture multi-scale features with excellent performance. However, compared with other models, there is still a certain computational overhead. Based on the Swin-T structure, UPerNet can better deal with large-scale and long-distance dependent information, and the mIoU and accuracy performance is good, but the computational complexity is high, resulting in relatively slow inference speed, a large parameter number and computational effort. SETR PUP is based on Vision Transformer Tiny (ViT-T) and fully adopts Transformer architecture. It can efficiently model global context information, but it may be slightly insufficient in local feature processing, resulting in slightly lower mIoU and accuracy. Segmenter is based on ViT-S and is suitable for tasks requiring global information modeling, but has a low mIoU in pitting image segmentation. SwinUnet is based on Swin Transformer, which can efficiently process global and local information of images, but it requires a large amount of computation and has a relatively slow inference speed. Segformer uses the MiT-B0 architecture, which is suitable for operation in the case of limited computing resources, and can maintain good accuracy while reducing computational complexity. Compared to SegFormer, the proposed method improves mIoU by 2.4% and reduces Flops by 54.72% on the DSBata test set, achieving an excellent balance between computational efficiency and segmentation accuracy. The experimental results confirm the effectiveness of the proposed method in the metal corrosion segmentation task.

In summary, the model proposed performs excellently in metal corrosion image segmentation tasks. It can correctly classify corrosion types, accurately identify corrosion boundaries, and effectively segment small corrosion areas, significantly reducing the probabilities of false positives and false negatives in the corrosion region. The proposed model has also demonstrated outstanding performance in the self-built pipeline corrosion pit image segmentation task and the BSData pitting image segmentation task. Compared to traditional segmentation methods, the proposed approach can accurately recognize the shape and size of corrosion pits, significantly reducing the issues of false negatives and false positives. Especially in handling complex and variable corrosion pit boundaries, the Edge-feature Extraction Module (EEM) effectively captures subtle edge features. Moreover, the introduction of the Feature Fusion Module (FFM) further enhances the model's ability to integrate multi-scale information and restore details, thereby improving the model's robustness and generalization ability.

Failure cases

However, despite the good performance of the proposed model in multiple tasks, the 67.69% mIoU on the metal surface corrosion image dataset is relatively low. As shown in Fig. 9, in the detection of large corrosion areas, the model tends to misclassify severe corrosion regions as poor or even fair corrosion, resulting in poor segmentation performance. On the one hand, this phenomenon is caused by the unbalanced distribution of datasets and the unbalanced classification of corroded images in metal surface corrosion public image dataset used in the

Sobel	Laplacian	mIoU↑	Acc↑	FPS↑	Params↓	Flops↓
	✓	90.78	99.73	66.49	3.87	7.16
✓		91.22	99.75	67.66	3.87	7.07
✓	✓	91.11	99.75	64.19	3.88	7.24

Table 10. Impact of multi-scale convolution.

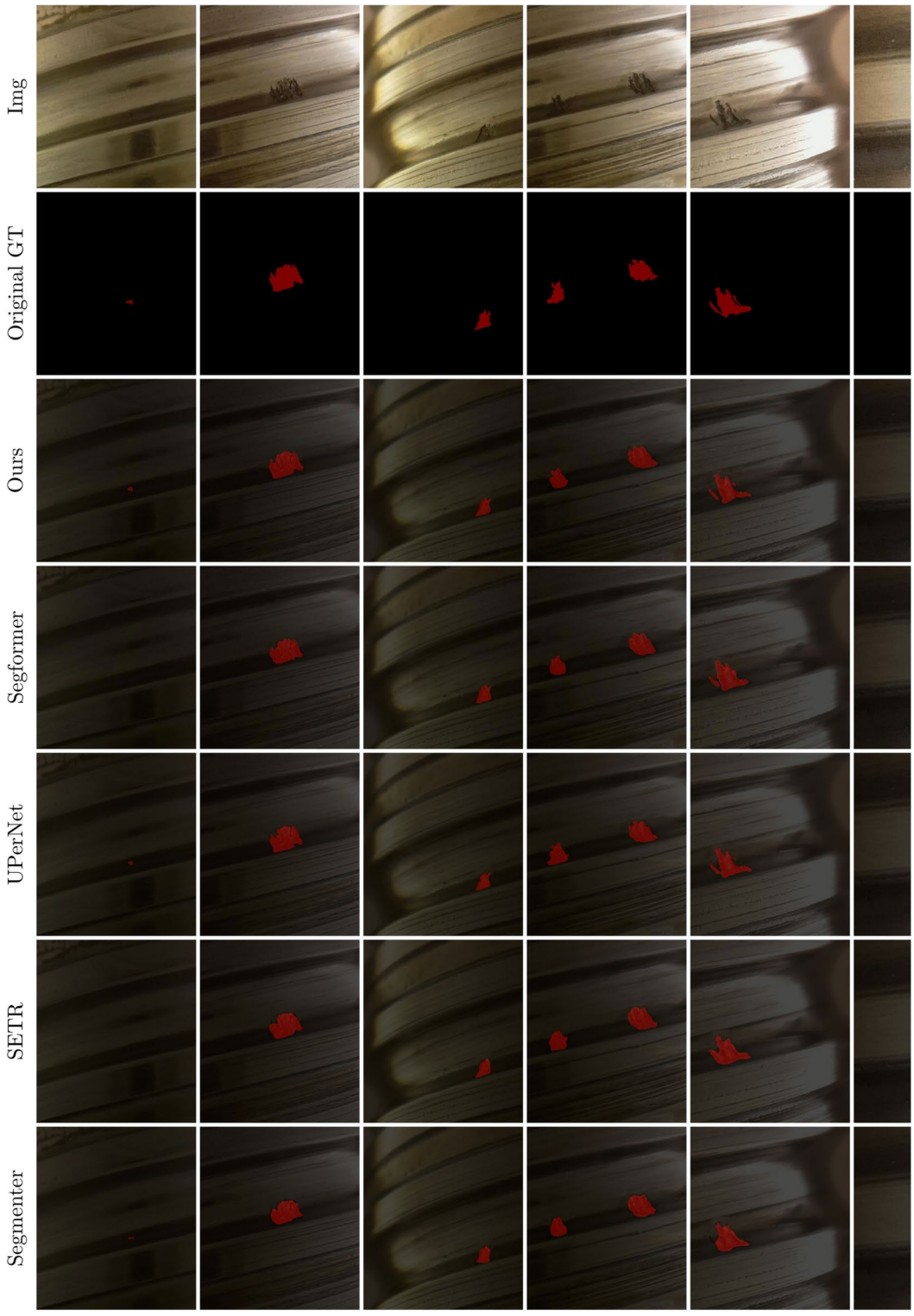

Fig. 8. Visualization results of different semantic segmentation models on BSData dataset.

experiment. In the dataset, the proportion of fair corrosion and poor corrosion is more, and the proportion of severe corrosion is less, which leads to the tendency of the model to predict more mild corrosion and moderate corrosion during the training process, so the model performs poorly in the segmentation of severely corroded regions. The other is the complexity of the task itself. Not only does the model need to accurately segment corroded areas, but it must also accurately distinguish between fairly, poorly, and severely corroded areas, which greatly increases the difficulty of the task. In contrast, the public dataset BSData and self-built tubing corrosion pit image dataset only require accurate segmentation of the corrosion pit or pitting region, which is a relatively

Model	Backbone	mIoU↑	Acc↑	FPS↑	Params↓	Flops↓
U-Net[50]	VGG-16	86.88	99.59	20.67	24.89	172.93
HRNet[52]	HRNetV2-W18	89.04	99.67	23.51	9.64	14.29
DeepLabv3+[53]	MobileNetV2	89.86	99.69	67.87	5.81	26.43
UPerNet[54]	Swin-T	90.18	99.74	15.42	58.94	179
SETR PUP[38]	ViT-T	89.13	99.68	5.32	308	309
Segmenter[42]	ViT-S	88.54	99.66	13.78	25.89	26.60
TransUnet[40]	ViT-B	88.35	99.61	25.60	93.23	32.24
SwinUnet[55]	Swin-T	88.89	99.60	24.69	41.34	34.79
Segformer[46]	MiT-B0	89.01	99.67	**75.37**	3.72	5.19
Ours	MiT-B0	**91.41**	**99.75**	71.02	**3.60**	**2.35**

Table 11. Evaluation results of different semantic segmentation models on BSData dataset. Significant values are in bold.

Fig. 9. Failed cases in the experiment.

single task. Under this condition, the proposed model can perform segmentation accurately, and ensure that the segmentation boundary is clear and correct.

Limitations and future work

(1) The relatively small number of data samples provided in both the public data set and our self-built corrosion data set limits the generalization ability and performance of the model. To improve the performance of the model, future work will focus on building larger corrosion datasets to ensure that the model can be trained and tested in a more diverse set of scenarios, thereby improving its overall performance.
(2) We do not discuss the performance of the model in the case of extreme noise or highly reflective surfaces. In some practical application scenarios, there may be strong background noise or highly reflective surfaces in the etched area, which will adversely affect the segmentation effect. Future work will focus on how to improve the robustness of models in these complex environments, possibly by introducing noise suppression techniques, reflection correction algorithms, or customizing model structures for specific environments to optimize model performance.

(3) In terms of data enhancement, although we have adopted an effective enhancement strategy, how to choose the most appropriate enhancement method according to the characteristics of different datasets still needs further research. Future work will focus on developing more intelligent and adaptive data augmentation techniques to improve the performance and adaptability of models in a variety of different scenarios.

Conclusion

To address the issue of poor segmentation performance in existing semantic segmentation algorithms for metal surface corrosion images, this paper proposed a SegFormer metal corrosion detection method based on the parallel extraction of edge features using the SegFormer architecture and the problem of how to accurately identify the target boundary and how to reduce the missing detection of small-area target regions. To solve the problem that the SegFormer network frequently uses the Transformer self-attention mechanism, which causes the network to pay too much attention to high-level abstract information and ignore shallow low-level information such as the edge and texture in the image, this paper used the edge-feature extraction module to build spatial branches of the network to extract image edge information. Aiming at the problem that the original network decoder is too simple, which leads to the loss of target feature information, this paper employs depthwise separable convolutions and feature concatenation operations to construct the feature fusion upsampling decoding layer. This approach enhances feature fusion while reducing detail loss, thereby improving the network's detection accuracy for small target areas and simultaneously decreasing the number of model parameters and computational complexity. The experimental results show that compared with the baseline SegFormer network, the proposed method improves 1.49% mIoU on the public metal surface corrosion image dataset. On the self-built tubing corrosion pit image dataset, the accuracy metrics of 89.32%mIoU are achieved with the parameter number of 3.60 MB. It reached 91.41% mIoU on the public BSData dataset. Compared with other networks, the proposed method has higher segmentation accuracy, fewer parameters and computational complexity. The significant advantages of the proposed method on different datasets validate its applicability and effectiveness in practical application scenarios.The lightweight characteristics of the proposed method further enhance its applicability and effectiveness in resource-constrained scenarios.

Data availability

The metal surface corrosion image dataset used in the current study can be downloaded from the following link: https://data.lib.vt.edu/. The BSData dataset used in the current study can be downloaded from the following link: https://github.com/2Obe/BSData. The self-built tubing corrosion pit image dataset used in the current study is not publicly available due to its current confidential status, but is available from the corresponding author on reasonable request.

Received: 30 October 2024; Accepted: 28 February 2025
Published online: 08 March 2025

References

1. Mazumder, M. A. J. Global impact of corrosion: Occurrence, cost and mitigation. *Glob. J. Eng. Sci.* **5**(4), 4. https://doi.org/10.33552/GJES.2020.05.000618 (2020).
2. Bender, R. et al. Corrosion challenges towards a sustainable society. *Mater. Corros.* **73**(11), 1730–1751. https://doi.org/10.1002/maco.202213140 (2022).
3. Vorobel, R., Ivasenko, I., Berehulyak, O. & Mandzii, T. Segmentation of rust defects on painted steel surfaces by intelligent image analysis. *Autom. Constr.* **123**, 103515. https://doi.org/10.1016/j.autcon.2020.103515 (2021).
4. Ma, S., Song, K., Niu, M., Tian, H. & Yan, Y. Cross-scale fusion and domain adversarial network for generalizable rail surface defect segmentation on unseen datasets. *J. Intell. Manuf.* **35**(1), 367–386. https://doi.org/10.1007/s10845-022-02051-7 (2024).
5. Shah, S. & Tembhurne, J. Object detection using convolutional neural networks and transformer-based models: A review. *J. Electr. Syst. Inf. Technol.* **10**, 54. https://doi.org/10.1186/s43067-023-00123-z (2023).
6. Zhao, H., Jia, J. & Koltun, V. *Exploring Self-Attention for Image Recognition* 2004–13621. https://doi.org/10.48550/arXiv.2004.13621 (2020).
7. Parmar, N. et al. *Image Transformer*, vol. 80, 4055–4064. https://doi.org/10.48550/arXiv.1802.05751 (2018).
8. Safa, A., Mohamed, A., Issam, B. & Mohamed-Yassine, H. Segformer: Semantic segmentation based tranformers for corrosion detection. In *2023 International Conference on Networking and Advanced Systems (ICNAS)* 1–6. https://doi.org/10.1109/ICNAS59892.2023.10330461 (2023).
9. Tang, K., Zhang, P., Zhao, Y. & Zhong, Z. Deep learning-based semantic segmentation for morphological fractography. *Eng. Fract. Mech.* **303**, 110149. https://doi.org/10.1016/j.engfracmech.2024.110149 (2024).
10. Fukuoka, T. & Fujiu, M. Detection of bridge damages by image processing using the deep learning transformer model. *Buildings* **13**(3), 788. https://doi.org/10.3390/buildings13030788 (2023).
11. Zhang, C., Cui, J., Wu, J. & Zhang, X. Attention mechanism and texture contextual information for steel plate defects detection. *J. Intell. Manuf.* **35**, 2193–2214. https://doi.org/10.1007/s10845-023-02149-6 (2024).
12. Tian, J., Zeng, Z., Hong, Z. & Zhen, D. Research on salient object detection algorithm for complex electrical components. *J. Intell. Manuf.* https://doi.org/10.1007/s10845-024-02434-y (2024).
13. Pedram, M. et al. Objective characterisation of reinforced concrete with progressive corrosion defects through clustering and thresholding of infrared images. *Measurement* **225**, 114017. https://doi.org/10.1016/j.measurement.2023.114017 (2024).
14. Palakal, M. J., Pidaparti, R. M., Rebbapragada, S. & Jones, C. R. Intelligent computational methods for corrosion damage assessment. *AIAA J.* **39**, 1936–1943. https://doi.org/10.2514/2.1183 (2001).
15. Dong, B. et al. Monitoring reinforcement corrosion and corrosion-induced cracking by X-ray microcomputed tomography method. *Cem. Concr. Res.* **100**, 311–321. https://doi.org/10.1016/j.cemconres.2017.07.009 (2017).
16. Choi, K. Y. & Kim, S. S. Morphological analysis and classification of types of surface corrosion damage by digital image processing. *Corros. Sci.* **47**(1), 1–15. https://doi.org/10.1016/j.corsci.2004.05.007 (2005).
17. Dorafshan, S., Thomas, R. J. & Maguire, M. Comparison of deep convolutional neural networks and edge detectors for image-based crack detection in concrete. *Constr. Build. Mater.* **186**, 1031–1045. https://doi.org/10.1016/j.conbuildmat.2018.08.011 (2018).
18. Diaz, J. A. I., Ligeralde, M. I., Jose, J. A. C. & Bandala, A. A. *Rust Detection Using Image Processing Via Matlab 2017–December* 1327–1331. https://doi.org/10.1109/TENCON.2017.8228063 (2017).

19. Sun, L., Li, Y., Li, X. & Liu, C. Corrosion defect segmentation method based on superpixel feature cascade. *Ain Shams Eng. J.* **15**(2), 102425. https://doi.org/10.1016/j.asej.2023.102425 (2024).
20. Ahuja, S. K. & Shukla, M. K. A survey of computer vision based corrosion detection approaches. *Inf. Commun. Technol. Intell. Syst.* **84**, 55–63. https://doi.org/10.1007/978-3-319-63645-0_6 (2018).
21. O'Byrne, M., Schoefs, F., Ghosh, B. & Pakrashi, V. Texture analysis based damage detection of ageing infrastructural elements. *Comput. Aided Civil Infrastruct. Eng.* **28**, 1. https://doi.org/10.1111/j.1467-8667.2012.00790.x (2013).
22. Liu, L., Tan, E., Zhen, Y., Yin, X. J. & Cai, Z. Q. *AI-Facilitated Coating Corrosion Assessment System for Productivity Enhancement* 606–610. https://doi.org/10.1109/ICIEA.2018.8397787 (2018).
23. Srivastava, A., Ji, G. & Singh, R. K. *Application of Deep-Learning Architecture for Image Analysis Based Corrosion Detection* 1–5. https://doi.org/10.1109/STCR51658.2021.9588887 (2021).
24. Fondevik, S. K., Stahl, A., Transeth, A. A. & Knudsen, O. Ø. *Image Segmentation of Corrosion Damages in Industrial Inspections* 787–792. https://doi.org/10.1109/ICTAI50040.2020.00125 (2020).
25. Tan, L., Chen, X., Yuan, D. & Tang, T. Dsnet: A computer vision-based detection and corrosion segmentation network for corroded bolt detection in tunnel. *Struct. Control Health Monit.* https://doi.org/10.1155/2024/1898088 (2024).
26. Huang, J. et al. A lightweight residual model for corrosion segmentation with local contextual information. *Appl. Sci.* **12**(18), 95. https://doi.org/10.3390/app12189095 (2022).
27. Yin, B. et al. *Deepsc-Edge: Scientific Corrosion Segmentation with Edge-Guided and Class-Balanced Losses* 1662–1668. https://doi.org/10.1109/ICMLA58977.2023.00251 (2023).
28. Xiao, X. et al. Baseg: Boundary aware semantic segmentation for autonomous driving. *Neural Netw.* **157**, 460–470. https://doi.org/10.1016/j.neunet.2022.10.034 (2023).
29. Zhu, Z. et al. Brain tumor segmentation based on the fusion of deep semantics and edge information in multimodal mri. *Inf. Fusion* **91**, 376–387. https://doi.org/10.1016/j.inffus.2022.10.022 (2023).
30. Jagadeesh, B. & Anand Kumar, G. Brain tumor segmentation with missing mri modalities using edge aware discriminative feature fusion based transformer u-net. *Appl. Soft Comput.* **161**, 111709. https://doi.org/10.1016/j.asoc.2024.111709 (2024).
31. Tsai, T.-H. & Tseng, Y.-W. Bisenet v3: Bilateral segmentation network with coordinate attention for real-time semantic segmentation. *Neurocomputing* **532**, 33–42. https://doi.org/10.1016/j.neucom.2023.02.025 (2023).
32. Bui, N.-T., Hoang, D.-H., Nguyen, Q.-T., Tran, M.-T. & Le, N. *Meganet: Multi-scale Edge-Guided Attention Network for Weak Boundary Polyp Segmentation* 7970–7979. https://doi.org/10.1109/WACV57701.2024.00780 (2024).
33. Ma, P. et al. A laplace operator-based active contour model with improved image edge detection performance. *Dig. Signal Process.* **151**, 104550. https://doi.org/10.1016/j.dsp.2024.104550 (2024).
34. Li, X., Yu, S., Gu, H., Tan, Y. & Xing, L. Underwater image clearing algorithm based on the Laplacian edge detection operator. In *Proceedings of 2nd International Conference on Artificial Intelligence, Robotics, and Communication* (eds Yadav, S. et al.) 159–172 (Springer, 2023)
35. Guo, B., Zhang, J. & Li, X. River extraction method of remote sensing image based on edge feature fusion. *IEEE Access* **11**, 73340–73351. https://doi.org/10.1109/ACCESS.2023.3296641 (2023).
36. Biehl, M., Ghosh, A. & Hammer, B. Dynamics and generalization ability of lvq algorithms. *J. Mach. Learn. Res.* **8**, 323–360. https://doi.org/10.1007/s10846-006-9096-7 (2007).
37. Li, X. et al. Semantic segmentation of uav remote sensing images based on edge feature fusing and multi-level upsampling integrated with deeplabv3+. *PLoS ONE* **18**, 9097. https://doi.org/10.1371/journal.pone.0279097 (2023).
38. Zheng, S. et al. *Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers* 2012–15840. https://doi.org/10.48550/arXiv.2012.15840 (2020).
39. Carion, N. et al. *End-to-End Object Detection with Transformers* 213–229. http://arxiv.org/abs/2005.12872 (2020).
40. Chen, J. et al. *Transunet: Transformers Make Strong Encoders for Medical Image Segmentation*. http://arxiv.org/abs/2102.04306v1 (2021).
41. Dosovitskiy, A. et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale* 2010–11929. https://doi.org/10.48550/arXiv.2010.11929 (2020).
42. Strudel, R., Garcia, R., Laptev, I. & Schmid, C. *Segmenter: Transformer for Semantic Segmentation* 7262–7272. http://arxiv.org/abs/2105.05633 (2021).
43. Han, K., Guo, J., Tang, Y. & Wang, Y. *Pyramidtnt: Improved Transformer-in-Transformer Baselines with Pyramid Architecture*. http://arxiv.org/abs/2201.00978 (2022).
44. Wu, S., Wu, T., Lin, F., Tian, S. & Guo, G. *Fully Transformer Networks for Semantic Image Segmentation*. https://doi.org/10.48550/arXiv.2106.04108 (2021).
45. Wang, W. et al. *Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction Without Convolutions* 548–558. https://doi.org/10.1109/ICCV48922.2021.00061 (2021).
46. Xie, E. et al. *Segformer: Simple and Efficient Design for Semantic Segmentation with Transformers*, vol. 34, 12077–12090. http://arxiv.org/abs/2105.15203 (2021).
47. Bianchi, E. & Hebdon, M. *Corrosion Condition State Semantic Segmentation Dataset* (2021).
48. Schlagenhauf, T., Landwehr, M. & Fleischer, J. Industrial machine tool component surface defect dataset. *CoRR*. http://arxiv.org/abs/2103.13003 (2021).
49. Yan, H., Zhang, C. & Wu, M. Lawin transformer: Improving semantic segmentation transformer with multi-scale representations via large window attention. *CoRR*. http://arxiv.org/abs/2201.01615 (2022).
50. Ronneberger, O., Fischer, P. & Brox, T.: U-net: Convolutional networks for biomedical image segmentation. *CoRR*. http://arxiv.org/abs/1505.04597 (2015).
51. Long, J., Shelhamer, E. & Darrell, T. Fully convolutional networks for semantic segmentation. *CoRR*. http://arxiv.org/abs/1411.4038 (2014).
52. Sun, K., Xiao, B., Liu, D. & Wang, J. Deep high-resolution representation learning for human pose estimation. *CoRR*. http://arxiv.org/abs/1902.09212 (2019).
53. Chen, L., Zhu, Y., Papandreou, G., Schroff, F. & Adam, H.: Encoder–decoder with atrous separable convolution for semantic image segmentation. *CoRR*. http://arxiv.org/abs/1802.02611 (2018).
54. Xiao, T., Liu, Y., Zhou, B., Jiang, Y. & Sun, J. Unified perceptual parsing for scene understanding. *CoRR*. http://arxiv.org/abs/1807.10221 (2018).
55. Cao, H. et al. *Swin-unet: Unet-Like Pure Transformer for Medical Image Segmentation*. http://arxiv.org/abs/2105.05537 (2021).

Author contributions

Bingnan Yan: Writing—Review & Editing, Funding acquisition. Conghui Wang: Conceptualization, Writing—Review & Editing. Xiaolong Hao: Writing—Review & Editing, Funding acquisition.

Funding

This research was supported by the National Natural Science Foundation of China (41904112), the Natural Science Basic Research Program of Shaanxi Province (2024JC-YBMS-201), and the Graduate Innovation and

Practical Ability Training Program of Xi'an Shiyou University (YCS23213118).

Declarations

Competing interests
The authors declare no competing interests.

Consent for publication
All authors have reviewed the manuscript and approved it for submission.

Additional information
Correspondence and requests for materials should be addressed to X.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.