# Corrosion SAM: Adapting Segment Anything Model with Parameter-Efficient Fine-Tuning for Structural Corrosion Inspection

Chengzhang Chai, Yan Gao, Haijiang Li*, Xiaofeng Zhu
BIM for Smart Engineering Centre, School of Engineering, Cardiff University, UK
chaic1@cardiff.ac.uk

**Abstract.** Corrosion is one of the primary factors leading to structural deterioration, necessitating regular inspections to maintain. Existing instrument-based inspections are prohibitively expensive, while computer vision-based models face obstacles in generality and adaptability across different scenes. This study introduces a novel framework that incorporates a Parametric Efficient Fine-Tuning (PEFT) strategy into the vision foundation model, the Segmentation Anything Model (SAM), to improve the task of structural corrosion inspection. Our approach bridges the gap between the SAM pre-trained model and the downstream task by transferring knowledge from the natural image domain to the structural corrosion domain. The PEFT strategy significantly reduces the consumption of computational resources and shortens the model's training time while maintaining high performance on new tasks. The effectiveness and superiority of the proposed approach have been verified through a series of comparative experiments conducted on two structural corrosion datasets, as well as the potential of the foundation model.

## 1. Introduction

Corrosion is one of the common structural defects and a significant cause of failure. Therefore, it is advisable to conduct regular inspections to maintain the infrastructure structure. Recently, there have been three main types of corrosion inspection methods. The first method is to complete the inspection with the help of physical testing instruments such as ultrasonic detectors, x-ray equipment, and electrochemical testing equipment (Vasagar et al., 2024). Although these devices can get very accurate inspection results, they often have the limitations of the high cost of instruments and the time-consuming and labour-intensive inspection process. The second category is using traditional image processing techniques (Ahuja and Shukla, 2018). It can effectively extract regional features in the image and thus identify the specifics of corrosion. This type of method has the advantages of low cost and simple processing. However, manual feature extraction will have time-consuming problems when facing complex environments. Additionally, this method exhibits poor sensitivity to texture and colour features, which can be a significant limitation in detecting corrosion types. In contrast, the third type of deep learning-based corrosion detection models can automatically learn high-dimensional abstract semantic features (Zhang et al., 2021), which can segment the corroded region at the pixel level. However, current defect detection models are insufficiently generalised and adaptable when crossing scenarios (Tulbure et al., 2022). When adapting to new tasks in different scenarios, a large amount of data is often required to complete the fine-tuning of the model. Such a process typically consumes extensive computational resources. Therefore, it is necessary to investigate more advanced deep learning models to perform the task of structural corrosion detection better.

Foundation models, also known as pre-trained large models, are considered the next generation of general paradigms in Artificial Intelligence (AI). Specifically, the family of Large Language Models (LLM), such as Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) and Generative Pre-trained Transformers (GPT) (Radford et al., 2019), along with others, have achieved great success in the field of Natural Language Processing (NLP). These models learn patterns and associations from vast amounts of data through deep

learning frameworks established by pre-training on extensive datasets. When generalised to various downstream tasks using techniques such as fine-tuning, they can capture rich a prior knowledge through powerful generic feature representation capabilities and have demonstrated superior comprehension in language-related tasks. Similarly, the foundation model is being extended to the field of vision. Segmentation Anything Model (SAM) was developed and trained on a dataset of 11 million images and 1.1 billion masks, and it demonstrated impressive visual perceptual segmentation performance (Kirillov et al., 2023). This pioneering research in image segmentation provides a valuable opportunity for the downstream task. However, since the pre-training dataset for the visual foundation model is primarily natural images, it does not perform well in several specific downstream tasks (Chen et al., 2023) (Ji et al., 2023). Therefore, it is worth exploring further how to utilise the segmentation capabilities of the visual foundation model in downstream tasks. This study will focus on the structural corrosion inspection task.

Given this, this paper proposes a novel framework that incorporates the Parameter Efficient Fine-Tuning (PEFT) strategy into the foundation model SAM for better application to structural corrosion inspection tasks. Our approach aims to bridge the gap between the SAM pre-trained model and the downstream task so that complex corrosion environments can be accurately identified and segmented. In addition, the PEFT strategy allows knowledge migration from the natural image domain to the structural corrosion domain by adjusting only a tiny portion of the model parameters. This strategy not only drastically reduces the consumption of computational resources but also shortens the model training time while maintaining the model's high performance on new tasks. We validate the effectiveness and superiority of the proposed approach through extensive experiments on two structural corrosion datasets. This research advances the application of SAM to specific downstream tasks and provides a new perspective and methodology for solving field-specific problems using foundation models.

The subsequent sections of this paper are organised as follows: Section 2 reviews related work on structural corrosion detection and the foundation model SAM. Section 3 details the architectural design of the Corrosion SAM we developed. Section 4 outlines the experimental design and implementation steps, presenting the results of ablation studies and comparative experiments. Finally, Section 5 summarises the research and main findings in this paper while providing an outlook on potential directions for future research.

## 2. Related work

### 2.1 Structural corrosion inspection

In recent years, computer vision-based corrosion detection methods have become mainstream, especially in the industrial and infrastructure fields. These methods mainly utilise advanced deep learning models, such as Convolutional Neural Network (CNN) and Transformer models, which have shown great potential in corrosion detection. Cha et al. (2018) developed a structural damage detection method based on the Faster R-CNN model for detecting damages, including steel corrosion and bolt corrosion. He demonstrated promising results and accuracy on a database containing 2366 images. Atha and Jahanshahi (2018) applied two CNN-based networks, ZF Net and VGG, to a database of contrasting images of 33,039 corroded and 34,148 uncorroded regions. A fine-tuning approach was used to demonstrate the CNN network's state-of-the-art. Forkan et al. (2022) proposed an integrated framework for identifying and detecting corrosion called CorrDetector based on CNN models. Excellent and effective results were obtained when evaluated on complex structural images collected in infrastructure

environments. Wang and Su (2023) proposed two deep learning models, BearDet and BearCla, using the Transformer architecture as a feature extraction network. The ability to detect and classify the corrosion level of bridge bearings, a critical structural component, was demonstrated in experiments.

Most previous work has used traditional CNN models or the Transformer architecture. However, CNN models have significant limitations in global modelling in handling dependencies due to small receptive fields and spatial invariance, which is likely to result in a limited model performance. Furthermore, the Transformer architecture relies on the self-attention mechanism to automatically capture global dependencies, providing a more robust framework to understand long-distance image interactions. Nevertheless, the Transformer architecture also faces new challenges, as fine-tuning requires the use of large-scale datasets to learn complex patterns. It also means that when the amount of data is not large enough, there is a risk of overfitting the model when fine-tuning the parameters, which poses a challenge to the generalisation and adaptability of the model. On the other hand, the full-param fine-tuning of the model requires sufficient computational resources to provide support. Therefore, it is necessary to use more efficient fine-tuning strategies to investigate more advanced deep-learning models.

### 2.2 Foundation model SAM

The SAM marks an essential innovation in foundation models in vision. With its unique composition of image encoder, prompt encoder, and mask decoder, pre-training on large-scale datasets enables it to recognise various visual elements ranging from simple textures to complex scenes. In medical image segmentation, the MedSAM model developed by Ma et al. (2024) demonstrated its generalised segmentation capability on different medical image datasets by fully fine-tuning the SAM. However, its reliance on high-end computational resources (20 GPUs of NVIDIA A100 80GB) limited its general application. To address this issue, Gong et al. (2023) proposed a new adaptive technique that significantly improves the performance of medical image segmentation tasks through the fusion of lightweight adapters and domain-specific knowledge. In remote sensing image segmentation, Pu et al. (2024) also demonstrated that by fine-tuning some parameters, the SAM model can accomplish efficient land cover segmentation with low computational resource requirements.

Although the SAM model has performed well within several fields, such as medicine and remote sensing, it outperforms many traditional supervised learning models. However, it still faces considerable challenges in specific application scenarios, such as structural corrosion detection. Structural corrosion is characterised by its tiny, irregular damage regions that are not only difficult to distinguish from the background but also have a high degree of uncertainty in the morphology and distribution of the damage. These features require models that are not only highly sensitive but also can adapt to complex and changing backgrounds and corrosion types. Therefore, how to better accomplish the task of structural corrosion inspection with the help of foundation models has become an issue worth exploring.

## 3. Methodology

### 3.1 Overall framework

As mentioned, the foundation model SAM should be optimised when migrating to new downstream tasks. The standard architecture of SAM is shown in the upper part of Fig. 1. After inputting an image into the image encoder, the model encodes the image through a patch

embedding process and several Transformer layers and finally outputs it to a mask decoder that generates the corresponding output mask. However, when faced with a scenario in an unseen domain, such as structural corrosion, the standard structure of SAM cannot adequately capture the specific features of the corrosion image. This is because its pre-training weights are mainly aimed at the natural image domain rather than the structural corrosion image domain. As a result, even if an output mask can be generated through the Transformer layer and mask decoder, the accuracy and effectiveness of the mask will be significantly reduced. As shown in the output of the upper part of Fig. 1, corrosion regions cannot be correctly extracted.

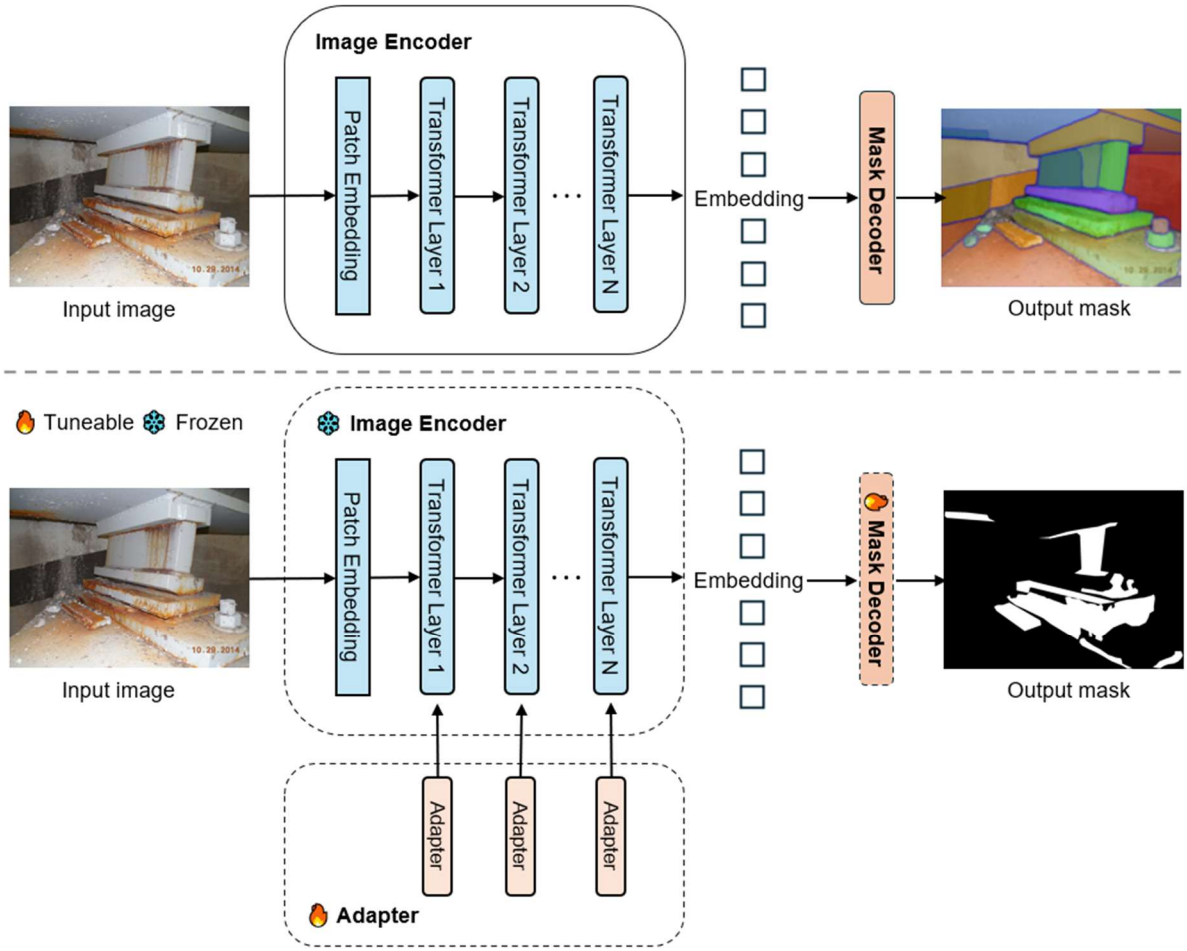


Figure 1: The proposed overall framework (Upper: Original SAM, Lower: Corrosion SAM)

Given this, a model architecture with an adapter called the corrosion SAM is proposed. Its overall framework is shown in the lower part of Fig. 1. In this architecture, most Transformer layers are frozen and no longer updated during the fine-tuning process. Instead, only the adapter and mask decoder parameters are updated. The adapters work in each Transformer layer and are responsible for fine-tuning the model to a specific downstream task based on the feature representation of the Transformer layer.

### 3.2 SAM Transformer architecture

The standard backbone network of SAM uses Vision Transformer (ViT) architecture, as shown in the left part of Fig. 2. This architecture captures the long-range dependencies of the input image through its multi-head attention mechanism to efficiently encode complex local areas. Each Transformer layer contains a multi-head attention unit that extracts information from multiple patch levels in parallel and ensures numerical stability during training by layer

normalisation. The next multilayer perceptron (MLP) further processes the features to enhance the nonlinear representation of the model. Crucially, the residual connectivity between layers allows some inputs to be passed directly to deeper layers. This approach facilitates deeper feature learning, prevents the gradient vanishing problem, and ensures efficient training of deeper networks. Together, the above designs form an efficient Transformer module.

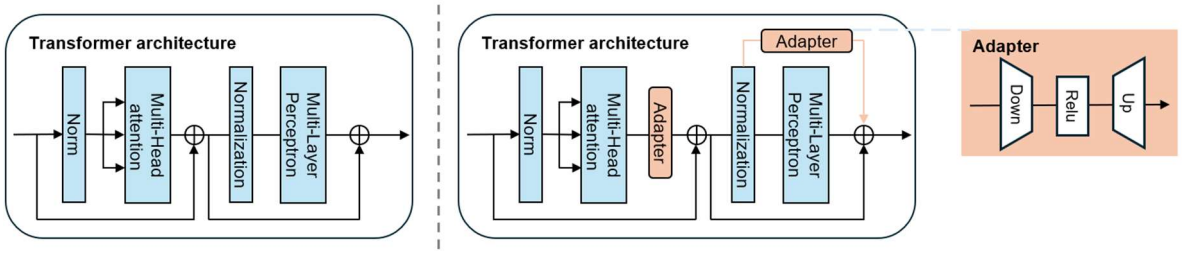


Figure 2: The Transformer layer architecture (Left: Original ViT, Right: Improved ViT)

Taking the standard ViT architecture as a basis, we strategically embedded adapter modules into the Transformer layer, as shown in the right part of Figure 2. These modules are placed after the multi-head attention module and parallel to the MLP block. We are introducing an adapter after the multi-head attention module, which allows efficient task fine-tuning with only a few trainable parameters on top of the original pre-trained model weights. At the same time, placing the adapter parallel to the MLP block achieves another level of adaptation in the feature extraction process. Such a setup allows feature transformations in different representation spaces, improving the model's sensitivity to specific tasks and maintaining enough flexibility to learn new feature representations.

### 3.3 Adapter module

The internal structure of the adapter is shown in the orange area on the right side of Fig. 2. It consists of a set of lightweight layers that project activations into a low-dimensional space, process them through a nonlinear activation function, and then project the activations back to the original dimension. This "compression-activation-expansion" model allows the Adapter module to learn task-relevant feature transformations at a much lower parameter cost. It does not interfere with the rich representations captured by the pre-trained network. This process is more cost-effective and faster regarding computational resources than full-param fine-tuning.

## 4. Experiment validation

### 4.1 Dataset preparation

We conducted validation experiments on the proposed method using two open-source structural corrosion datasets, the VDOT dataset (Bianchi and Hebdon, 2023) and the KRDB dataset (Fujishima et al., 2023), to demonstrate the method's effectiveness.

a. VDOT dataset

The VDOT dataset is derived from bridge inspection reports managed by the Virginia Department of Transportation (VDOT). This dataset was annotated strictly following the corrosion condition guidelines of the Bridge Inspector's Reference Manual (BIRM) and the American Association of State Highway and Transportation Officials (AASHTO). The original dataset consisted of 440 finely annotated images in four corrosion condition classes (good, fair, poor, and severe). Considering that the dataset has a significant category imbalance, i.e., there

are far fewer annotations in the poor and severe classes than in the fair classes. We chose to merge the annotations in the poor and severe classes with the annotations in the fair classes so that the merged dataset will be segmented only for corrosion regions in the structural image. This merge will reduce the disturbance caused by category imbalance to the model on the one hand and reduce the computational complexity of the model on the other hand.

b. KRDB dataset

The KRDB dataset is derived from bridge inspection reports managed by the Kanto Regional Development Bureau (KRDB). This dataset contains five categories of damage: corrosion, cracking, free lime, water leakage, and spalling, each finely annotated. We selected only the structural corrosion images from it, totalling 955 images.

For each dataset, 80%, 10%, and 10% are selected as training, validation, and test sets, respectively. To further improve the model performance, we uniformly adopted the normalisation process of binarization for the image pixel values of the dataset and the operation of resizing the image to 512 × 512.

### 4.2 Model evaluation metrics

The model's performance is evaluated using Intersection over Union (IoU) and Dice coefficient, commonly used evaluation metrics in segmentation tasks (Wang et al., 2022). These metrics are intended to measure the degree of overlap and approximation between the masks segmented by the model and the ground truth.

a. IOU

IoU is a metric used to evaluate the performance of image segmentation algorithms. It represents the ratio of intersection and concatenation between the predicted segmentation result and the actual region. The formula is calculated as follows, and the value of IoU ranges from 0 to 1. The closer its value is to 1, the higher the overlap between the predicted result, and the ground truth and the higher the similarity.

$$IoU(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

b. Dice coefficient

The Dice coefficient is a pixel-level metric for assessing the similarity of two samples. It is usually used to compare the similarity between segmentation results and ground truth. It is calculated as follows: twice the intersection part divided by the sum of the respective sizes of the two samples. The values range from 0 to 1, where 0 indicates no overlap and 1 indicates complete overlap.

$$Dice(A,B) = \frac{2|A \cap B|}{|A| + |B|} \tag{2}$$

### 4.3 Implementation details

All experiments were executed in Python 3.8 and Pytorch 1.13.0 environments with an Intel(R) Core (TM) i9-12900 CPU@ 2.40 GHz processor and 128 GB of RAM. It also has an NVIDIA RTX A6000 to ensure high computational efficiency and robustness. Considering that we are comparing the method developed in this paper with other existing segmentation algorithms, the implementation parameters of all the algorithms were kept as consistent as possible when conducting the experiments.

The training process is configured for a maximum of 100 epochs. The image size accepted by the model is 512×512, and the input batch size is 8. Considering the large number of irregular shapes and significant differences in image quality in the structural corrosion segmentation task, it may pose a challenge to the training and generalisation ability of the model. Therefore, on the one hand, the loss function selects the strategy of Cross entropy loss and Dice loss. Cross entropy loss is adept at dealing with pixel-level classification issues and can effectively drive the model to learn to differentiate between corroded and non-corroded regions. Dice loss strengthens the model's ability to capture irregular shapes by quantifying the degree of overlap between predicted and actual regions.

On the other hand, these neural networks use the 'Adam optimiser' and 'Step decay' learning rate schedule strategy. This is because this combined strategy can flexibly adapt to different training stages. Initially, the 'Adam optimiser' adaptive ability quickly adapts to the data's irregular shape and quality changes, ensuring that the model can converge quickly in complex environments. As the training progresses, the 'step decay' learning rate adjustment strategy reduces the overfitting risk of the model in the later stages of training by gradually decreasing the learning rate. Simultaneously, it refines the model's ability to capture details and enhances its generalisation. The initial learning rate is then set to 1e-4. Specific implementation details are shown in the following Table 1.

Table 1: Fine-tuning implementation parameter

| Parameter | Corrosion SAM (Ours) | Existing segmentation method |
|---|---|---|
| Epochs | 100 | 100 |
| Image size | 512×512 | 512×512 |
| Batch size | 8 | 8 |
| Loss function | Cross entropy loss & Dice loss | Cross entropy loss & Dice loss |
| Optimizer type | Adam | Adam |
| Learning rate schedule | Step decay | Step decay |
| Initial learning rate | 1e-4 | 1e-4 |

### 4.4 Experimental results

To validate the model, we conducted ablation studies and comparison experiments on the VDOT and KRDB datasets. In the ablation study section, we obtained the best model performance by comparing different model sizes of the backbone. In the comparison experiments section, we select a variety of existing methods that perform well and compare them to the Corrosion SAM to assess the overall performance of our method. Finally, we then show the effect of model segmentation through visualisation.

a. Ablation study

The backbone of the ViT, which is pivotal to our model's enhancement, is available in three distinct sizes: ViT-H, ViT-L, and ViT-B. Experiments conducted on the VDOT dataset indicate that adjusting the backbone from ViT-B to ViT-H results in a marked performance improvement. Specifically, the IoU metric increased from 64.01% to 69.72%, and the Dice score increased from 76.33% to 81.04%. The KRDB dataset shows a similar pattern of change. This increasing trend suggests that larger model sizes can lead to better performance, likely due to their increased capacity for capturing more complex patterns within the data. Therefore, we chose the better-performing ViT-H as the backbone of our model.

Table 2: Ablation study

| Backbone | VDOT dataset | | KRDB dataset | |
|---|---|---|---|---|
| | IoU (%) | Dice (%) | IoU (%) | Dice (%) |
| ViT-B | 64.01 | 76.33 | 55.93 | 68.53 |
| ViT-L | 67.12 | 78.85 | 57.17 | 70.08 |
| ViT-H | **69.72** | **81.04** | **60.78** | **72.97** |

b. Comparison experiment

The existing algorithms compared in this section involve DeepLabV3+ (Chen et al., 2018), U-Net (Ronneberger et al., 2015), and SegFormer (Xie et al., 2021). Each method has a well-designed structure. DeepLabV3+ combines a null convolution and an encoder-decoder structure for capturing multi-scale features and improving the accuracy of segmented edges. U-Net uses a symmetric encoder-decoder structure featuring jump connections between the encoder and decoder to improve the segmentation of small targets. SegFormer is a lightweight and efficient Transformer-based semantic segmentation model that improves performance by combining multi-scale features with a self-attention mechanism. When conducting the comparison experiments, we set the same implementation parameters for the above method as Corrosion SAM, as detailed in Table 1. The comparison results are shown in Table 3.

Table 3: Comparison of with other methods

| Method | Backbone | VDOT dataset | | KRDB dataset | |
|---|---|---|---|---|---|
| | | IoU (%) | Dice (%) | IoU (%) | Dice (%) |
| DeepLabV3+ | MobileNet-V2 | 62.23 | 67.15 | 57.85 | 64.25 |
| U-Net | VGG-16 | 64.98 | 71.26 | 51.62 | 63.35 |
| SegFormer | MiT-B0 | 68.71 | 72.33 | 57.93 | 67.16 |
| Corrosion SAM | ViT-H | **69.72** | **81.04** | **60.78** | **72.97** |

As seen from the table, on both dataset, VDOT and KRDB, the Corrosion SAM achieves the best IoU and Dice coefficient evaluation metrics, which are superior to the other methods. Analysing the reasons, while CNN-based methods like DeepLabV3+ and U-Net have shown commendable performance in many image segmentation tasks, they may be limited by their local receptive fields when dealing with specific tasks that require an understanding of a broader context and capturing finer texture details. Meanwhile, SegFormer, although better able to capture global information through the Transformer architecture, may be deficient in feature recognition and context understanding capabilities as it has not been pre-trained on large-scale datasets. In contrast, the Corrosion SAM, which uses ViT-H as the backbone network, takes advantage of the robust learning capabilities during the pre-training phase of the foundation model, resulting in a more pronounced advantage in the integration of global information, complexity recognition of features, and task-specific context sensitivity. Additionally, its adapter module effectively narrows the domain gap between the foundational model trained on natural images and the specific demands of structural corrosion imagery, thereby enhancing its precision in identifying crucial features. These factors enable our model to achieve better results in structural corrosion detection tasks.

c. Visualisation results

Fig. 4 presents the visualisation results of the Corrosion SAM compared with other methods. Corrosion SAM can extract features of corroded regions more accurately and comprehensively, especially details such as edges and boundaries. When dealing with corroded edges and complex textures, other methods are more prone to over-segment and incorrectly classify non-corroded regions as corroded areas, affecting the evaluation's accuracy. In addition, they also often miss detecting corrosion features, especially if the corrosion is mild or the features are not significant enough in the image, and this under-segmentation may lead to misjudgement of structural safety. Corrosion SAM somewhat remedies these shortcomings by adapting foundation models to the structural corrosion domain using an adapter strategy. However, its segmentation results in certain fine-grained regions also suffering from ambiguous noise. This may be because the model still has limitations in dealing with highly similar textures or colour gradients and does not have enough capabilities to resolve such subtle differences. The next research phase could consider introducing more advanced multi-scale analysis and contextual enhancement strategies to improve the model's feature extraction capabilities in these challenging regions.

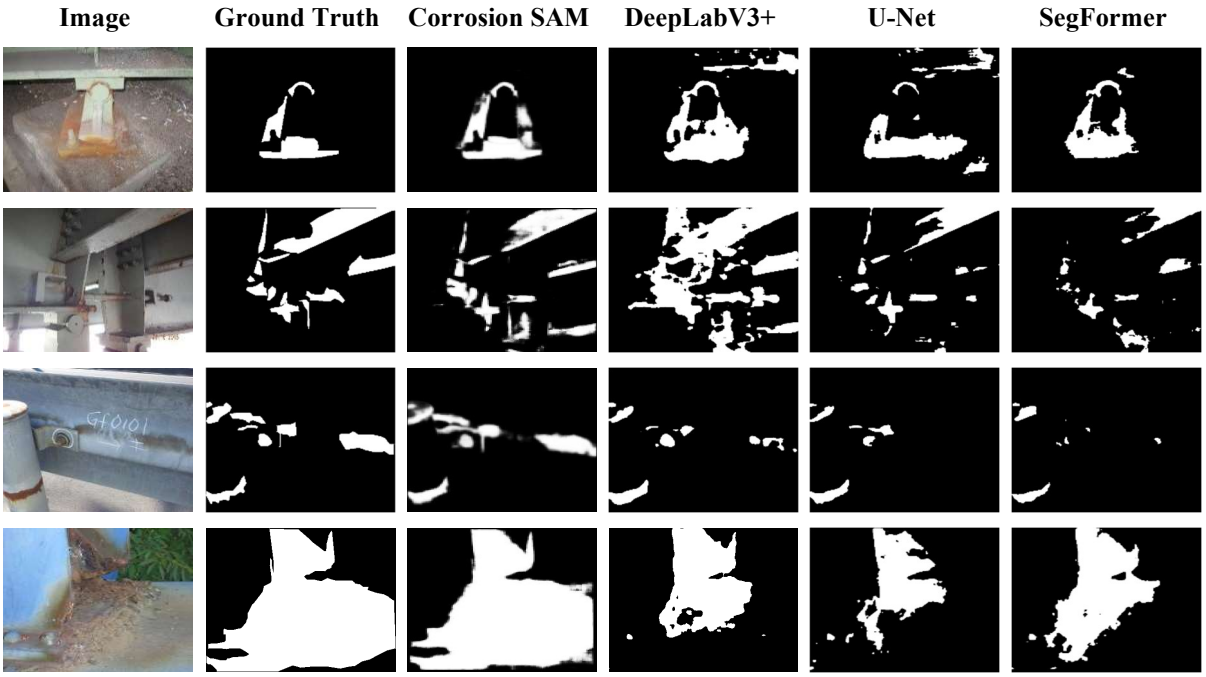


Figure 4: Visualisation results of different methods

## 5. Conclusions

This study presents a Corrosion SAM approach for inspecting structural corrosion. It incorporates PEFT strategies into the foundation model SAM, which can bridge the domain gap between natural and corrosion scenarios. Our proposed method demonstrates superior performance compared to existing approaches through comparative experiments on the VDOT and KRDB datasets, coupled with visualisation results. These findings affirm the effectiveness of our method and illuminate the potential application of visual foundation models in structure inspection.

In summary, this study offers a valuable contribution to structural corrosion inspection within the infrastructure, exploring the feasibility of applying foundation models to more actual

engineering fields. Future research can further improve the model's robustness and generalisation capabilities and explore more types of structural defects, paving the way for AI-driven structural inspection.

## References


Ahuja, S.K. and Shukla, M.K. (2018). A survey of computer vision based corrosion detection approaches. Information and Communication Technology for Intelligent Systems (ICTIS 2017)-Volume 2 2, pp.55-63.

Atha, D.J. and Jahanshahi, M.R. (2018). Evaluation of deep learning approaches based on convolutional neural networks for corrosion detection. Structural Health Monitoring, 17(5), pp.1110-1128.

Bianchi, E. and Hebdon, M. (2023). Available at: https://doi.org/10.7294/16624663.v2

Cha, Y., Choi, W., Suh, G., Mahmoudkhani, S., Büyüköztürk, O. (2018). Autonomous Structural Visual Inspection Using Region-Based Deep Learning for Detecting Multiple Damage Types. Computer-Aided Civil and Infrastructure Engineering, 33(9), pp. 731–747.

Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F. and Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European conference on computer vision (ECCV) (pp. 801-818).

Chen, T., Zhu, L., Ding, C., Cao, R., Zhang, S., Wang, Y., Li, Z., Sun, L., Mao, P. and Zang, Y. (2023). SAM fails to segment anything? --SAM-Adapter: Adapting SAM in underperformed scenes: Camouflage, shadow, and more. arXiv preprint arXiv:2304.09148.

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Forkan, A.R.M., Kang, Y.B., Jayaraman, P.P., Liao, K., Kaul, R., Morgan, G., Ranjan, R. and Sinha, S. (2022). CorrDetector: A framework for structural corrosion detection from drone images using ensemble deep learning. Expert Systems with Applications, 193, p.116461.

Fujishima, T., Dang, J. and Chun, P. (2023), Available at: https://doi.org/10.50915/data.jsceiii.24750210.v1

Gong, S., Zhong, Y., Ma, W., Li, J., Wang, Z., Zhang, J., Heng, P.A. and Dou, Q. (2023). 3dsam-adapter: Holistic adaptation of sam from 2d to 3d for promptable medical image segmentation. arXiv preprint arXiv:2306.13465.

Ji, W., Li, J., Bi, Q., Li, W. and Cheng, L. (2023). Segment anything is not always perfect: An investigation of sam on different real-world applications. arXiv preprint arXiv:2304.05750.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y. and Dollár, P. (2023). Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 4015-4026).

Ma, J., He, Y., Li, F., Han, L., You, C. and Wang, B. (2024). Segment anything in medical images. Nature Communications, 15(1), p.654.

Pu, X., Jia, H., Zheng, L., Wang, F. and Xu, F. (2024). ClassWise-SAM-Adapter: Parameter Efficient Fine-tuning Adapts Segment Anything to SAR Domain for Semantic Segmentation. arXiv preprint arXiv:2401.02326.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), p.9.

Ronneberger, O., Fischer, P. and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18 (pp. 234-241).

Tulbure, A.A., Tulbure, A.A. and Dulf, E.H. (2022). A review on modern defect detection models using DCNNs–Deep convolutional neural networks. Journal of Advanced Research, 35, pp.33-48.

Vasagar, V., Hassan, M.K., Abdullah, A.M., Karre, A.V., Chen, B., Kim, K., Al-Qahtani, N. and Cai, T. (2024). Non-destructive techniques for corrosion detection: A review. Corrosion Engineering, Science and Technology, p.1478422X241229621.

Wang, W. and Su, C. (2023). Deep learning-based detection and condition classification of bridge steel bearings. Automation in Construction, 156, p.105085.

Wang, Y., Ahsan, U., Li, H. and Hagen, M. (2022). A comprehensive review of modern object segmentation approaches. Foundations and Trends® in Computer Graphics and Vision, 13(2-3), pp.111-283.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M. and Luo, P. (2021). SegFormer: Simple and efficient design for semantic segmentation with transformers. Advances in neural information processing systems, 34, pp.12077-12090.

Zhang, S., Deng, X., Lu, Y., Hong, S., Kong, Z., Peng, Y. and Luo, Y. (2021). A channel attention based deep neural network for automatic metallic corrosion detection. Journal of Building Engineering, 42, p.103046.