electronics

MDPI

Article

A Corrosion Segmentation Method for Substation Equipment Based on Improved TransU-Net and Multimodal Feature Fusion

Hailong Guo [1], Guangqi Lu [1,*], Jiuyu Guo [2], Zhixin Li [3], Xuan Wang [1] and Zhenbing Zhao [4]

[1] State Grid Gansu Electric Power Company Tianshui Power Supply Company, Tianshui 741000, China; guohailongtsgd@163.com (H.G.); 18419194995@163.com (X.W.)
[2] State Grid Gansu Electric Power Company, Lanzhou 730070, China; jiuyu0207@126.com
[3] Electric Power Research Institute of State Grid Gansu Electric Power Company, Lanzhou 730070, China; ayuan719@foxmail.com
[4] School of Electrical and Electronic Engineering, North China Electric Power University, Baoding 071003, China
* Correspondence: luguangqineepu@163.com; Tel.: +86-17733380738

Abstract

Substation equipment operating in harsh environments is highly susceptible to corrosion, yet conventional image segmentation methods often fail to achieve precise delineation of corroded regions. Here, we propose an enhanced TransU-Net-based approach for corrosion segmentation. Deformable convolution is incorporated into the encoder to strengthen the model's capacity to represent irregular corrosion morphologies. A composite color–texture fusion module is developed to jointly exploit color information from HSV and Lab spaces together with multi-scale texture features. In addition, a Shape-IoU loss function is introduced to refine boundary fitting and improve contour accuracy. Experimental evaluations demonstrate that the proposed method consistently outperforms state-of-the-art models across multiple metrics, achieving an Intersection over Union (IoU) of 75.42% and a Recall (PA) of 83.14%. These results confirm that the model substantially enhances corrosion recognition accuracy and edge integrity under complex background conditions, offering a promising strategy for intelligent maintenance of substation infrastructure.

Keywords: power vision; corrosion segmentation; TransU-Net; deformable convolution; color–texture fusion

check for updates

Academic Editor: Chiman Kwan

Received: 26 October 2025
Revised: 22 November 2025
Accepted: 24 November 2025
Published: 28 November 2025

Citation: Guo, H.; Lu, G.; Guo, J.; Li, Z.; Wang, X.; Zhao, Z. A Corrosion Segmentation Method for Substation Equipment Based on Improved TransU-Net and Multimodal Feature Fusion. *Electronics* **2025**, *14*, 4688. https://doi.org/10.3390/electronics14234688

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Power grid construction has always been a very important basic industry in economic and social development [1]. Metal components such as main transformers, circuit breakers, and transmission fittings in substations are frequently exposed to outdoor environments characterized by high humidity and salt spray, making them particularly vulnerable to corrosion. According to statistics, the annual corrosion rate of substation equipment in coastal regions exceeds 35%, with severe corrosion accounting for approximately 12%. Traditional manual inspections of a medium-sized substation typically require 8–12 h and depend heavily on subjective visual assessments, which can lead to oversight and inconsistent results [2]. Therefore, the development of intelligent corrosion detection methods is crucial for advancing the digital transformation of power grid maintenance.

Image-based corrosion detection techniques can be broadly categorized into traditional and deep learning approaches. Traditional methods primarily rely on handcrafted features. For instance, an improved pyramid evolutionary strategy (IPES)–based multithreshold

segmentation method was developed for rail fastener corrosion detection, achieving a peak signal-to-noise ratio (PSNR) of 28–35 dB, but exhibiting limited adaptability to diverse power equipment structures [3]. Other studies have quantified surface roughness using gray-level co-occurrence matrices (GLCM) combined with HSV color histograms to localize corrosion on steel structures; however, these methods remain sensitive to lighting variations [4]. A parameterized Retinex algorithm was introduced to suppress shadow interference, though its accuracy declined significantly on red-coated surfaces [5]. Additionally, a Fourier frequency–domain approach combined with RGB thresholding improved corrosion detection on steel bridges but failed to account for interference from surface oil or dust [6].

Deep learning has addressed many limitations of handcrafted feature methods. In object detection, an improved YOLOv8s model (RD-YOLOv8s), which integrates efficient channel attention, micro-scale detection layers, and scaled intersection-over-union loss, achieved a mean average precision (mAP 0.5) of 93.4% when trained on a dataset of 9270 substation equipment images, improving recall by 11.5–20.9% over Faster R-CNN [2]. Similarly, an attention-augmented detection framework achieved a frame rate of 70 FPS for real-time inspection, though it did not address corrosion severity quantification [7]. For vibration damper corrosion on transmission lines, an enhanced YOLOv5 incorporating adaptive anchor boxes and Complete-IoU loss achieved an mAP of 89.2%, effectively improving the detection of small corrosion targets [8]. Moreover, RGB-D image fusion networks that combine color and depth information reduced false positives in complex substation backgrounds, achieving an accuracy of 90.5% [9].

In the field of semantic segmentation, a modified U-Net with a MobileNetV3_Large encoder and depthwise separable residual convolution reduced the number of parameters by 81.18% while attaining an IoU of 77.43% [10]. A fully convolutional network (FCN) approach produced effective corrosion detection on reinforced concrete surfaces but lacked boundary optimization [11]. U-Net–based architectures have also been extended with attention mechanisms to improve segmentation accuracy for corrosion detection. For instance, hybrid attention–based U-Net models achieved promising results but suffered from increased model complexity and slower inference speeds [12]. Infrared–visible image fusion networks enabled graded corrosion assessment under challenging backlighting conditions [13]. The HEU-Net architecture utilized hybrid attention to improve mIoU to 85.5%, yet its adaptability to complex equipment backgrounds remained limited [14]. Transformer-based architectures have demonstrated strong potential in industrial defect segmentation, improving mIoU by 4.9–6.2% over convolutional baselines, albeit with higher computational complexity [15]. TransU-Net has demonstrated the Transformer's capacity for modeling long-range feature dependencies, offering new insights into corrosion segmentation [16]. Additionally, a semi-supervised extreme learning machine leveraging limited labeled and large unlabeled data effectively addressed the scarcity of corrosion samples, achieving 91.5% detection accuracy [17].

Despite these advances, challenges persist in corrosion segmentation for power equipment. Irregular corrosion patterns in substation equipment manifest in a variety of forms, such as pitting, flaking, and diffuse staining, often with unclear boundaries and inconsistent texture distributions. These characteristics lead to severe intra-class variation and class imbalance at the pixel level, making it difficult for traditional CNNs with fixed receptive fields and standard convolutions to generalize well. Moreover, variations in illumination and background clutter further reduce the reliability of color- or texture-based methods. Conventional segmentation networks often suffer from inaccurate boundary localization, weak representation of fine-grained features, and limited ability to adapt to spatial deformation of corrosion regions.

Recent research has attempted to address these issues through advanced feature interaction strategies. For example, Zhang et al. [18] proposed FCIHMRT, a feature cross-interaction hybrid module residual network, which enables multi-level semantic fusion by bridging encoder and decoder layers. Their design improves segmentation under challenging backgrounds by enhancing cross-scale and cross-depth feature flow. Inspired by this, our work introduces a novel fusion module—MCTAM—that extends this idea by incorporating multi-color space cues and directional texture features, enabling fine-grained representation of corrosion patterns in a lightweight and adaptive manner.

To address these challenges, this paper proposes a segmentation framework with the following innovations:

1. Deformable convolution is incorporated into the encoder to enhance spatial adaptability, enabling the network to dynamically sample features from corrosion regions with irregular shapes and boundary curvature.
2. A Multi-Color and Texture Analysis Module (MCTAM) is designed to fuse complementary color features from HSV and Lab color spaces with directional texture descriptors, thereby improving the model's ability to distinguish corrosion under varying illumination and complex backgrounds.
3. A novel Shape-IoU loss function is introduced to enforce boundary-aware optimization by integrating contour information into the loss formulation, which significantly improves segmentation accuracy and edge completeness.

These components are designed to complement each other, forming a unified approach capable of segmenting nonuniform corrosion under complex backgrounds and illumination scenarios.

2. Materials and Methods

2.1. Data and Method Overview

2.1.1. Image Acquisition and Annotation

The corrosion images used in this study were captured using industrial RGB cameras in substation environments. Each image has a resolution of 5184 × 3888 pixels, ensuring high spatial fidelity for detecting fine-grained corrosion textures. Corroded surfaces include painted metals, rusted bolts, and weathered equipment enclosures. The ground truth masks were annotated using the LabelMe 3.16.7 [19] tool, with polygon-based regions manually labeled by trained technicians. The annotations were subsequently converted into binary mask images in PNG format, where the corrosion region is marked as foreground (label 1) and the background as 0. No synthetic or augmented images were used in model training to maintain realism. An example of one of the images and its corresponding label is illustrated in Figure 1.

2.1.2. Overview of the Proposed Segmentation Network

This study proposes a corrosion segmentation network based on an improved TransU-Net architecture, as illustrated in Figure 1.

The network consists of two main components: an encoder and a decoder. The encoder adopts a hybrid structure combining convolutional neural networks (CNNs) and Transformer blocks to capture both local details and global contextual dependencies in a complementary manner. To accommodate the diverse morphologies of corrosion defects, deformable convolution (DC) is incorporated into the encoder to enhance spatial adaptability, thereby improving the network's ability to model irregular corrosion boundaries.

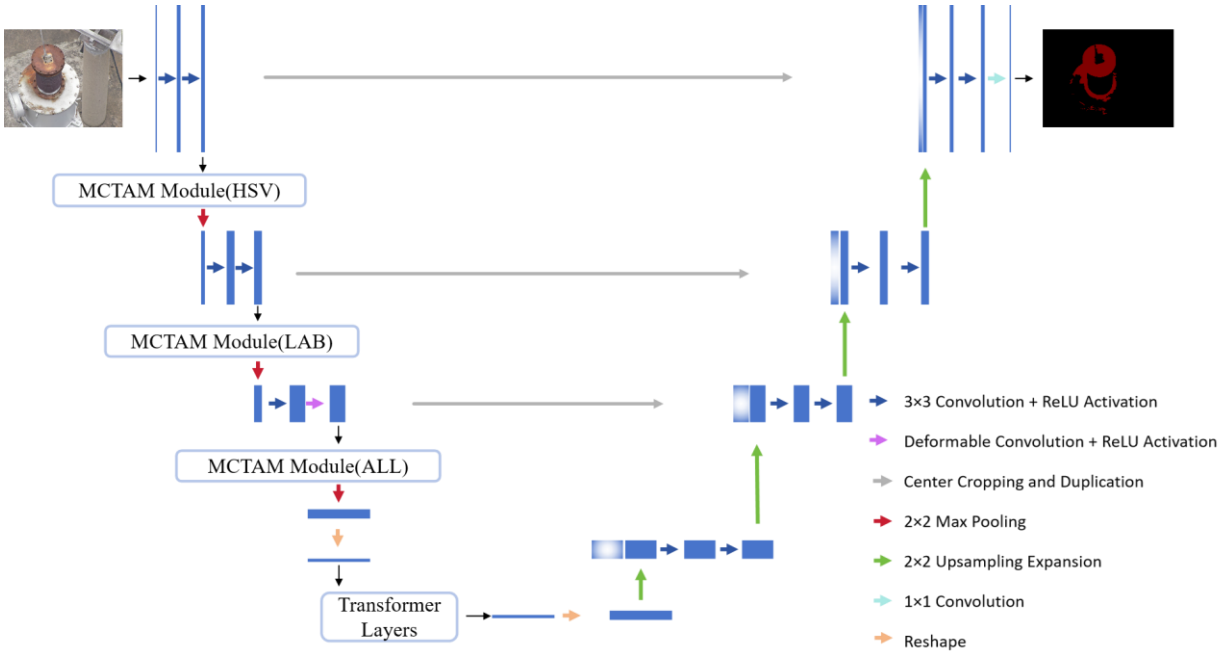


Figure 1. Overall architecture of the proposed corrosion segmentation network based on the improved TransU-Net.

Furthermore, during the down sampling process, a Multiple Color and Texture Analysis Module (MCTAM) is embedded at each stage. This module fuses 19-dimensional color and texture descriptors derived from both HSV and Lab color spaces, effectively enhancing corrosion feature representation. The multimodal fusion strategy has proven effective in various surface defect and corrosion detection tasks.

In addition, a novel Shape-IoU loss function is designed to jointly optimize region overlap and boundary shape consistency, thereby improving edge precision—particularly for small-scale corrosion regions and fuzzy boundaries. Through the integration of color–texture multimodal fusion and shape-aware optimization, the proposed network demonstrates robust performance and superior segmentation accuracy for substation equipment corrosion under complex background interference and blurred-edge conditions.

2.2. Multi-Color–Texture Analysis Module (MCTAM)

In substation equipment images, corrosion regions often exhibit subtle color variations, complex texture structures, and blurred boundaries. Their irregular morphology and diverse scales are frequently embedded within cluttered backgrounds, making it difficult for conventional models to effectively capture the critical features of corrosion. To address these challenges, a Multi-Color and Texture Analysis Module (MCTAM) is introduced and embedded at multiple stages of the encoder.

The MCTAM enhances the network's ability to extract corrosion-specific representations through color space transformation and multi-scale texture extraction. Its internal structure includes two parallel submodules: a color analysis path and a texture analysis path. The feature transformation process across channels—covering HSV and Lab color mapping, gradient computation, and frequency filtering—is illustrated in Figure 2, while the complete schematic of the MCTAM, showing its parallel fusion structure, is depicted in Figure 3.

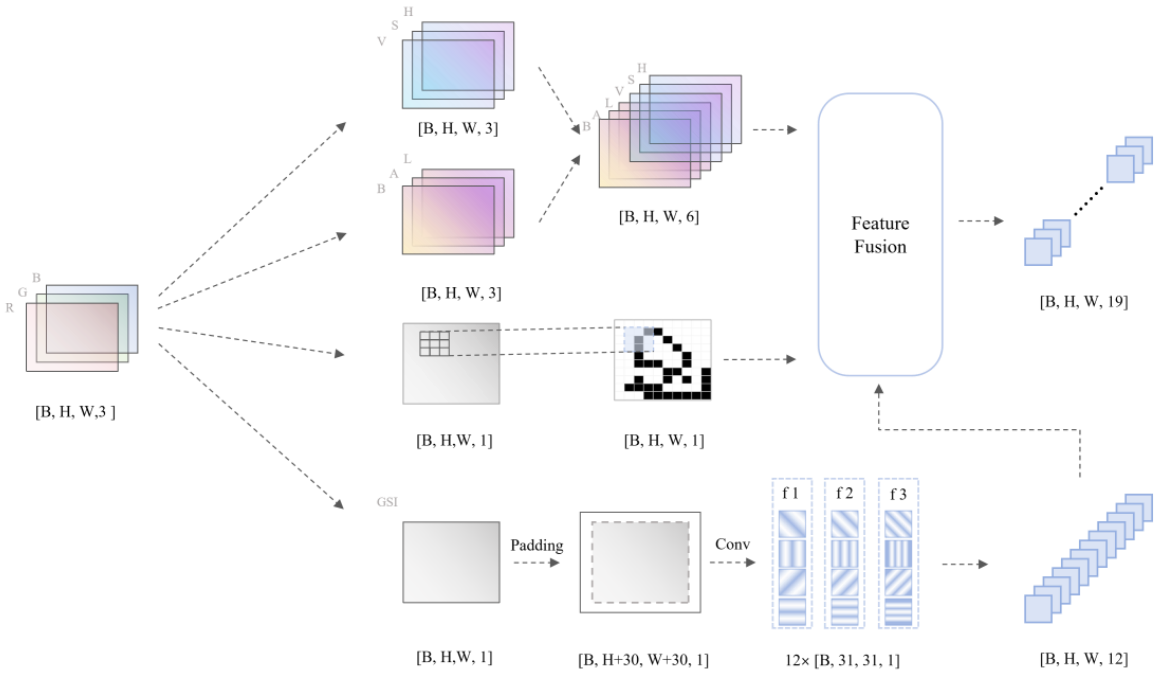


Figure 2. Channel-wise transformation process within MCTAM. This includes HSV and Lab color extraction, binary texture encoding, and directional frequency filtering.

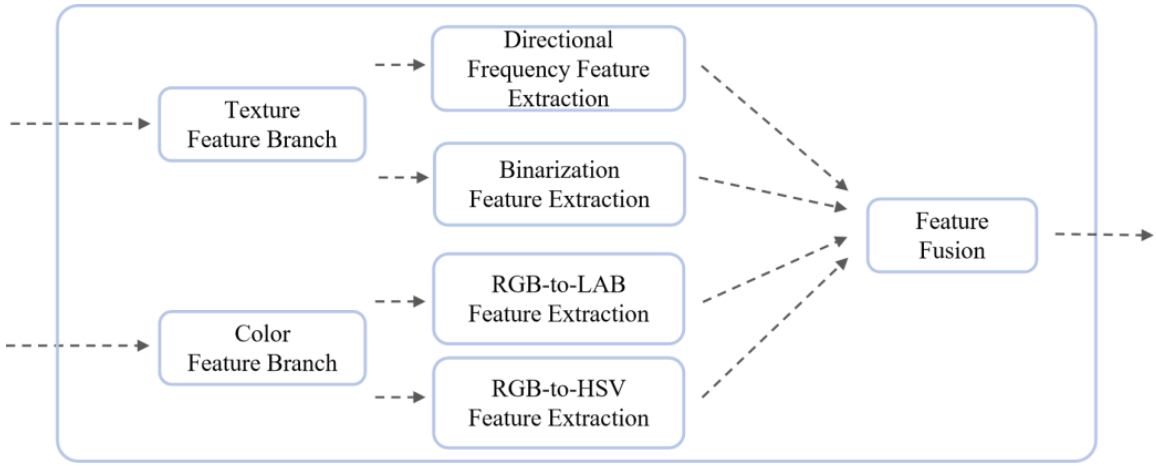


Figure 3. Schematic architecture of the Multi-Color and Texture Analysis Module (MCTAM), including parallel submodules for color analysis and texture analysis, followed by feature fusion.

To adaptively balance feature complexity and representation diversity across different network depths, the MCTAM is invoked at multiple encoder levels, with each invocation selectively enabling one or more sub-branches. Specifically, the terms "MCTAM (HSV)", "MCTAM (Lab)", and "MCTAM (ALL)" refer to the activation of only the HSV-based color features, only the Lab-based features, or all color and texture branches, respectively. These configurations are applied at different stages of the network and are not intended as ablation variants. This selective strategy allows the model to flexibly extract semantic information at varying scales while maintaining computational efficiency.

2.2.1. Color Analysis

The color analysis component takes the RGB image as input and maps it into two perceptually uniform color spaces: HSV and Lab. Three channels are extracted from each color space, yielding six color feature channels in total. These features strengthen the

model's capacity to identify subtle chromatic differences associated with corrosion. The HSV space effectively separates hue and luminance components, while the Lab space captures human perceptual sensitivity to color variation, improving robustness under varying illumination.

2.2.2. Texture Analysis

The texture analysis submodule operates at two levels. First, gray-level gradient computation is performed along eight orientations (0°–315°, with 45° increments) to capture local intensity variations. For each pixel, the gradient sum $S(i,j)$ is compared with those of its eight neighbors. Pixels with smaller gradient sums are assigned 0, and those with larger sums are assigned 1, forming a binary texture map. The computation is defined as:

$$g_k(i,j) = |G(i,j) - G(i+\Delta i_k, j+\Delta j_k)| (k = 1,2,...,8), \quad (1)$$

$$S(i,j) = \sum_{k=1}^{8} g_k(i,j), \quad (2)$$

$$(i,j) = \mathbb{I}(S(i+\Delta i_k, j+\Delta j_k) > S(i,j)), \quad (3)$$

where $\mathbb{I}(*)$ denotes the binary indicator function that returns 1 when the neighboring gradient sum exceeds that of the center pixel and 0 otherwise. $g_k(i,j)$ denotes the absolute gradient difference at pixel (i,j) in the k-th direction, measuring the intensity of gray-level variation between the central pixel and its neighboring pixel in that direction. $G(i,j)$ represents the gray value of the input grayscale image at pixel coordinate (i,j), where i is the row coordinate and j is the column Δi_k and Δj_k coordinate are the coordinate offsets in the k-th direction, corresponding to 8 directions. $S(i,j)$ is the sum of gradient differences at pixel (i,j) across all 8 directions, quantifying the overall texture roughness around the central pixel. This operation extracts a binary-level texture descriptor that captures directional intensity transitions.

Next, multi-frequency directional filtering is performed by combining Gaussian and sinusoidal functions to characterize spatial and frequency-domain texture variations. Convolution is applied using filters with frequencies of 0.1, 0.3, and 0.5, oriented at 0°, 45°, 90°, and 135°. The filters are expressed as:

$$f_{k,\theta}(x,y) = G(x,y)\cdot\sin(2\pi f_k\cdot(x\cos\theta + y\sin\theta)), \quad (4)$$

$$G(x,y) = \frac{1}{2\pi\sigma^2}\exp(-\frac{x^2+y^2}{2\sigma^2}), \quad (5)$$

$$F_{k,\theta}(i,j) = \sum_{x,y} I(x,y)\cdot f_{k,\theta}(i-x, j-y), \quad (6)$$

Here, $G(x,y)$ is a Gaussian function controlling spatial localization, and σ denotes its standard deviation. The convolution result $F_{k,\theta}(i,j)$ represents the filtered pixel value, where low-frequency filters (e.g., 0.1) emphasize coarse textures, while higher frequencies (e.g., 0.5) capture fine-grained textures. By combining three frequencies and four orientations, a total of 12 texture feature maps are generated, corresponding to multi-scale directional texture representations.

2.2.3. Feature Fusion

The one-dimensional gradient feature is concatenated with six color features and twelve frequency-domain texture features in the fusion layer, forming a 19-dimensional composite color–texture descriptor:

$$\mathrm{F_{MCTAM}} = \mathrm{Concat}(\mathrm{C_{HSV}}, \mathrm{C_{Lab}}, \mathrm{T_{grad}}, \mathrm{T_{filter}}), \quad (7)$$

where $C_{HSV} \in R$, $H \times W \times 3$ and $C_{Lab} \in R$, $H \times W \times 3$ represent the color features, $T_{grad} \in R$, $H \times W \times 1$ denotes the gradient-based texture feature, and $T_{filter} \in R$, $H \times W \times 1 \in R$ represents the multi-frequency texture responses. The fused representation $F_{MCTAM} \in R$, $H \times W \times 19$ is then propagated into the deeper layers of the main network to enhance discrimination of corrosion regions, particularly under complex backgrounds or blurred boundary conditions.

In MCTAM, feature fusion is implemented through a channel-wise concatenation followed by a 1×1 convolutional layer to reduce dimensionality and introduce learnable interaction among the combined features. Specifically, the outputs of the HSV color descriptor, Lab color descriptor, and texture branch are first normalized and then concatenated along the channel dimension. The resulting feature map is then passed through a 1×1 convolution followed by a batch normalization and ReLU activation, which enables the model to learn adaptive weighting among different types of features. This structure ensures efficient fusion while preserving discriminative information from each modality.

2.3. Deformable Convolution for Shape-Adaptive Feature Extraction

Conventional convolution operations employ fixed sampling positions, which limits their representational capacity when processing targets with irregular shapes and significant scale variations. In substation imagery, corrosion regions typically exhibit random spatial distribution, blurred boundaries, and cluttered backgrounds. Under such conditions, standard convolutions often extract irrelevant or misleading background information, thereby degrading segmentation accuracy.

To overcome these limitations, a deformable convolution (DC) structure is integrated into the final convolutional block of the encoder to enhance the shape adaptability of corrosion feature representation. Unlike standard convolutions with static sampling grids, deformable convolution introduces learnable spatial offsets that dynamically adjust the sampling locations within the receptive field. This mechanism enables the convolution kernel to align more closely with the geometric contours of corrosion regions.

The operation of a standard 2D convolution can be expressed as:

$$y(p_0) = \sum_{k=1}^{K} w(k) \cdot x(p_0 + p_k), \tag{8}$$

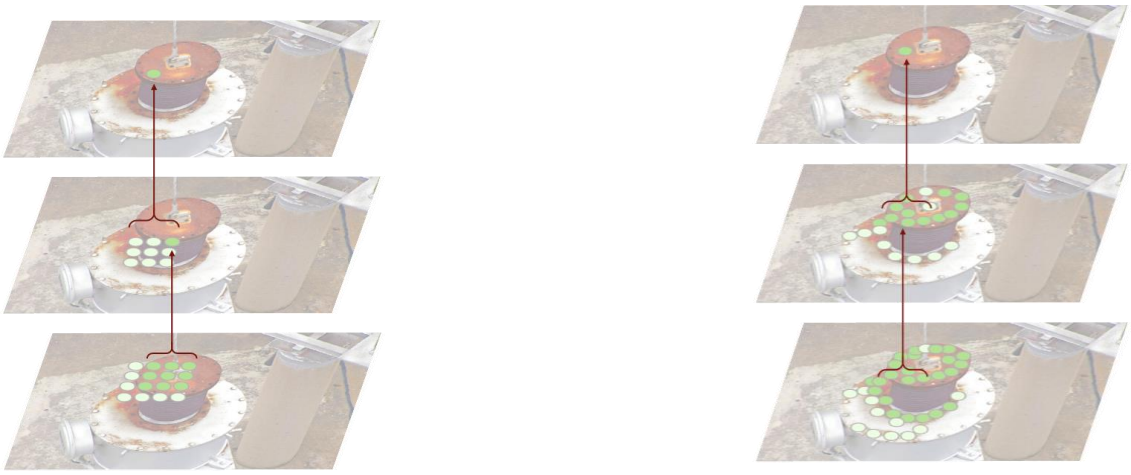
where $y(p_0)$ denotes the output feature value at position p_0, x is the input feature map, $w(k)$ represents the k-th kernel weight, and p_k is the fixed sampling location.

For deformable convolution, a learnable offset Δp_k is introduced to each sampling position, yielding:

$$y(p_0) = \sum_{k=1}^{K} w(k) \cdot x(p_0 + p_k + \Delta p_k), \tag{9}$$

The offset Δp_k is generated by an independent prediction branch and optimized through backpropagation. Consequently, each convolution kernel dynamically adapts its sampling pattern according to the local image structure, effectively accommodating geometric variations, boundary curvature, and fine-grained surface details within corrosion regions.

As illustrated in Figure 4, compared with the regular square sampling grid of standard convolution (Figure 4a), deformable convolution (Figure 4b) produces irregular sampling paths that better align with the actual target geometry. This allows the model to capture task-relevant features more precisely while significantly suppressing background interference.



(**a**) (**b**)

Figure 4. Comparison between standard convolution and deformable convolution in feature extraction. (**a**) Illustration of the Standard Convolution Operation; (**b**) Illustration of the Deformable Convolution Operation.

By integrating the deformable convolution module into the high-level feature extraction stage of the encoder, the proposed model achieves enhanced representation of corrosion regions with complex morphology, while maintaining the multi-scale modeling advantages of the original TransU-Net structure. This design ultimately improves the segmentation mask's accuracy and contour integrity.

2.4. Shape-IoU Loss for Boundary-Aware Optimization

In practical corrosion segmentation tasks, corrosion regions typically exhibit subtle inter-class variations, indistinct boundaries, and relatively small spatial proportions. These characteristics make conventional loss functions such as Cross-Entropy (CE) and Dice loss insensitive to boundary details during training, often resulting in discontinuous or incomplete target contours in the segmentation output. Such deficiencies can substantially degrade model performance in real-world inspection scenarios.

To enhance the model's capability to capture fine-grained shape and boundary structures, this study introduces a Shape-IoU loss function, designed to optimize the geometric fidelity of the segmentation mask. Unlike traditional overlap-based losses, the proposed loss simultaneously considers both the region-level similarity and the structural alignment of object contours, effectively guiding the model toward producing corrosion masks with sharper and more continuous boundaries.

The conventional Intersection-over-Union (IoU) loss is defined as

$$\mathcal{L}_{IoU} = 1 - \frac{|P \cap G|}{|P \cup G|}, \tag{10}$$

where denotes the predicted mask region, the ground-truth mask, and $|*|$ represents the number of pixels in a region. However, this formulation suffers from gradient vanishing issues when the overlap between P and G is small, leading to unstable training particularly for small corrosion targets.

To address this problem, the proposed Shape-IoU loss incorporates boundary contour information into the optimization objective. Specifically, the boundary sets ∂P and ∂G are

extracted from the predicted and ground-truth masks, respectively, and their intersection over union is computed as

$$\mathcal{L}_{Shape-IoU} = 1 - \frac{|\partial P \cap \partial G|}{|\partial P \cup \partial G| + \epsilon} \tag{11}$$

where ϵ is a small constant to prevent division by zero. Compared with conventional IoU, the Shape-IoU formulation emphasizes boundary conformity rather than regional overlap, making it particularly suitable for corrosion regions with fuzzy or irregular contours.

The pixel-wise cross-entropy loss $\mathrm{L}(CE)$ used in our framework is defined as

$$\mathrm{L}(CE) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{i,j} \times \log p_{i,j} \tag{12}$$

where N is the total number of pixels, C is the number of classes, $y_{i,j} \in \{0, 1\}$ indicates whether pixel i belongs to class c, and $p_{i,j}$ is the predicted probability for pixel i being assigned to class C. This formulation penalizes incorrect pixel-level class predictions and is effective for guiding region-level classification in segmentation tasks.

The overall loss function combines the standard pixel-wise cross-entropy loss $\mathcal{L}_{CE}$ and the proposed Shape-IoU loss as follows:

$$\mathcal{L}_{Total} = \lambda_{ce} \cdot \mathcal{L}_{CE} + \lambda_{shape} \cdot \mathcal{L}_{Shape-IoU}, \tag{13}$$

where λ_{ce} and λ_{shape} are weighting coefficients that balance pixel accuracy and structural consistency. In this study, the coefficients are set to $\lambda_{ce} = 1.0$ and $\mathcal{L}_{Shape-IoU} = 1.2$.

As illustrated in Figure 5, the Shape-IoU module effectively directs the network to focus on edge structural cues during training, mitigating the common issue of shape information neglect in conventional loss functions. Experimental results demonstrate that incorporating Shape-IoU significantly improves boundary recognition and segmentation consistency, particularly for fine-grained corrosion regions with complex textures and indistinct edges.

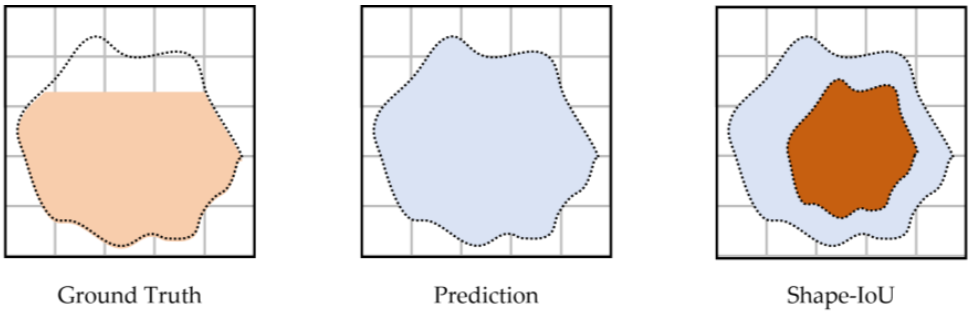


Figure 5. Schematic illustration of the Shape-IoU loss mechanism for boundary-aware optimization.

3. Experimental Results and Analysis

3.1. Experimental Data and Settings

In this study, corrosion defect images of substation equipment were used as the primary experimental dataset. All images were manually captured during routine substation inspections under real operating conditions. Each corrosion image was annotated at the pixel level using the professional labeling tool LabelMe. The annotation protocol followed these principles: regions where corrosion caused negligible color alteration were not labeled; for areas with blurred boundaries, a buffer treatment was applied to ensure smooth yet detailed edge annotation consistent with the actual corrosion morphology. Disconnected

corrosion regions were carefully refined to remove non-corrosive subareas, maintaining a consistent annotation style and ensuring high dataset reliability.

After preprocessing, a total of 515 annotated images were obtained, which were randomly divided into a training set (463 images) and a test set (52 images), corresponding to an approximate 9:1 ratio.

The experiments were conducted on an Ubuntu 20.04.2 operating system using an NVIDIA GTX 3090 GPU, with implementation based on the PyTorch 1.8.1 deep learning framework. Model training was performed for 200 epochs with a batch size of 8 and an initial learning rate of 0.0005, dynamically adjusted using the cosine annealing strategy. The AdamW optimizer was employed with a momentum of 0.9 and a weight decay coefficient of 0.01 to accelerate convergence and improve training stability.

3.2. Evaluation Metrics

To comprehensively assess the segmentation performance of the proposed model on substation corrosion images, three quantitative metrics were employed: Intersection over Union (IoU), Precision, and Recall (PA).

The IoU measures the overlap ratio between the predicted corrosion region and the ground-truth annotation. Precision represents the proportion of correctly predicted positive samples among all samples predicted as positive, while Recall (PA) quantifies the proportion of true positive samples that are correctly identified among all actual positive samples. The metrics are defined as follows:

$$IoU = \frac{TP}{FP + FN + TP} \tag{14}$$

$$Precision = \frac{TP}{TP + FP} \tag{15}$$

$$Recall = \frac{TP}{TP + FN} \tag{16}$$

where TP denotes true positives (correctly predicted corrosion pixels), FP represents false positives (non-corrosion pixels incorrectly predicted as corrosion), and FN denotes false negatives (corrosion pixels incorrectly predicted as non-corrosion).

3.3. Ablation and Visualization Experiment

To validate the effectiveness and individual contribution of each proposed component, a series of ablation experiments were conducted based on the baseline TransU-Net framework. The improved modules—Deformable Convolution (DC), Multiple Color and Texture Analysis Module (MCTAM), and Shape-IoU Loss—were introduced incrementally. The experimental results are summarized in Table 1.

Table 1. Ablation study results of the proposed corrosion segmentation network.

Baseline Network	DC	MCTAM	Shape-IoU	Mean Precision		Corrosion Precision		
				mIoU /%	mPA /%	IoU /%	PA /%	Precision /%
√				84.68	89.12	70.68	79.80	86.08
√		√		85.92	90.80	72.87	82.25	86.47
√		√	√	86.65	90.70	74.26	82.97	87.61
√			√	85.1	89.72	71.54	81.21	85.73
√	√	√	√	87.21	90.05	75.42	83.14	89.04

The "√" symbol indicates whether the experiment includes the selected module.

The observations can be summarized as follows:

The baseline TransU-Net achieved a PA of 79.80% and an IoU of 70.68%, indicating reasonable segmentation capability. However, due to the fixed sampling grid, a lack of explicit color–texture modeling, and weak boundary sensitivity, the model struggled with complex corrosion patterns, resulting in both false positives and missed detections.

Incorporating deformable convolution into the encoder improved mIoU to 85.92%, with IoU and PA increasing by 2.19% and 2.45%, respectively. This confirms that DC adaptively adjusts sampling positions to better conform to irregular corrosion geometries, effectively suppressing background interference and maintaining stable performance, especially in large, corroded areas or curved-edge regions.

With the addition of MCTAM, mIoU reached 86.65%, and Precision increased to 87.61%, representing a 1.53% gain over the baseline. The fusion of multimodal color (HSV/Lab) and multi-directional texture features compensates for the limitations of conventional convolution in capturing corrosion details, enhancing discriminative capability—particularly under weak contrast or uneven illumination—thus reducing omission of small corrosion areas.

Introducing the Shape-IoU loss alone raised IoU by 0.86% and PA by 1.41%, confirming its effectiveness in guiding the network to focus on edge structures. Although Precision slightly decreased (by 0.35%) due to a more inclusive boundary prediction, this trade-off improved contour continuity and completeness, which is advantageous for practical segmentation tasks.

When all three enhancements were integrated, the model achieved an mIoU of 87.21%, IoU of 75.42%, PA of 83.14%, and Precision of 89.04%, corresponding to respective improvements of 2.53%, 4.74%, 3.34%, and 2.96% over the baseline. These results demonstrate clear complementarity: DC strengthens geometric feature extraction, MCTAM enhances color–texture representation, and Shape-IoU Loss refines boundary optimization. Their combined use substantially improves robustness and generalization under complex substation environments.

Visualization results further confirm these findings. The baseline model often produced fragmented corrosion masks and discontinuous edges in fine corrosion spots and narrow cracks. With deformable convolution, corrosion contours became more complete; MCTAM enhanced recognition of low-contrast corrosion; and the combined model delivered smooth, connected, and visually coherent corrosion regions, closely matching human annotations. As shown in Figure 6, we selected three representative examples: (a) long strip-shaped and fine corrosion regions, (b) large-scale dispersed corrosion areas, and (c) edge-affected corrosion zones. Highlighted boxes indicate areas where the proposed model significantly reduced false positives and missed detections compared to the baseline.

In summary, the ablation study conclusively demonstrates that the three proposed modules are both effective and complementary. Their synergistic integration enables high-precision corrosion segmentation of substation equipment even under challenging background conditions.

3.4. Comparative Experiments

To comprehensively evaluate the performance of the proposed method, comparisons were conducted with six state-of-the-art semantic segmentation models, namely DeeplabV3+ [20], UNet, Segformer [21], SAM2 [22], TransU-Net, and Mask2Former [23]. In addition to accuracy metrics, Frames Per Second (FPS), Parameters (Params), and Floating Point Operations (FLOPs) were employed to assess computational efficiency. The quantitative results are summarized in Table 2.

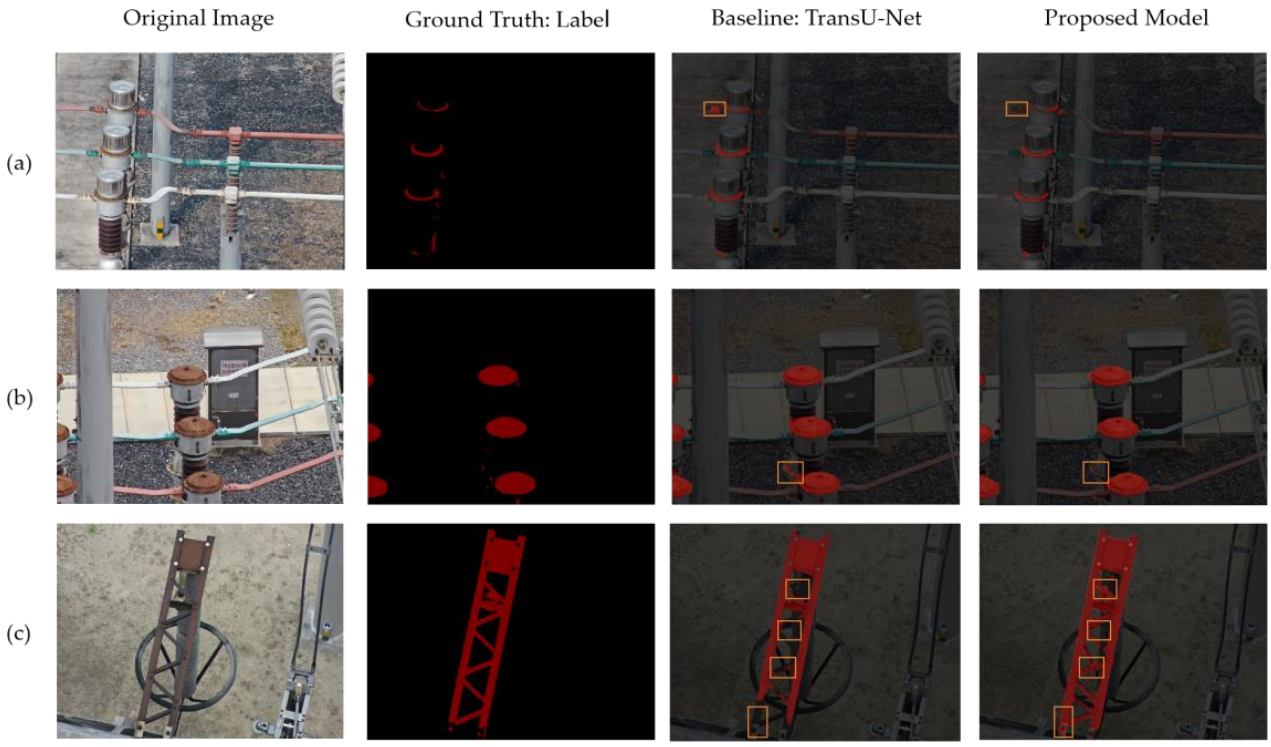


Figure 6. Visual comparison of segmentation results for representative corrosion scenarios: (**a**) strip-shaped and fine rust spots, (**b**) dispersed corrosion across entire surface, and (**c**) severe edge-interfered corrosion. From left to right: Original Image, Ground Truth, Baseline (TransU-Net), and Proposed Model. Highlighted boxes indicate reduced false detections and missed regions.

Table 2. Comparative analysis of the proposed method with mainstream semantic segmentation models.

Models	Mean Precision		Corrosion Precision			FPS	Params /M	FLOPs/T
	mIoU /%	mPA /%	IoU /%	PA /%	Precision /%			
DeeplabV3+	79.74	86.54	59.95	74.13	79.73	17.05	41.29	1.422
UNet	83.39	89.35	67.38	79.80	81.23	59.28	43.93	0.180
Segformer	85.46	90.65	71.41	82.33	84.34	31.02	27.34	0.055
SAM2	86.75	90.95	73.86	82.55	87.53	4.09	38.95	1.655
TransU-Net	84.68	89.12	70.68	79.80	86.08	29.00	105.28	0.016
Mask2former	84.18	88.86	69.69	79.16	85.35	18.46	46.74	0.216
Ours	87.21	90.05	75.42	83.14	89.03	26.56	108.73	0.014

From the comparative analysis, several observations can be drawn:

The convolutional model DeeplabV3+ achieved an mIoU of 79.74%, capturing multi-scale contextual information but showing limited adaptability to corrosion regions with fuzzy boundaries and irregular shapes due to its fixed convolutional kernels. UNet and TransU-Net, as typical encoder–decoder architectures, reached mIoU values of 83.39% and 84.68%, respectively. Both exhibited improved recall on small-scale corrosion regions compared with DeeplabV3+, yet struggled with false positives in complex backgrounds, indicating insufficient feature expressiveness.

The Segformer model, combining a lightweight Transformer encoder with a CNN decoder, achieved an mIoU of 85.46% with relatively low computational cost, demonstrating

that global dependency modeling significantly benefits the segmentation of irregular corrosion contours. SAM2, which incorporates image slicing and sequential modeling mechanisms, further improved performance, reaching an mIoU of 86.75% and a corrosion IoU of 73.86%, showing superior robustness in handling blurred boundaries and weak contrast areas—currently one of the best-performing general-purpose segmentation models.

Building upon these approaches, the proposed model integrates deformable convolution, the Multiple Color–Texture Analysis Module (MCTAM), and the Shape-IoU loss in a unified optimization framework. These components jointly enhance the model's ability to capture fine corrosion details and improve contour fitting. As a result, the model achieved an mIoU of 87.21% and a corrosion IoU of 75.42%, surpassing SAM2 by 0.46% and 1.56%, respectively. The Precision reached 89.04%, confirming the robustness of multimodal feature fusion for corrosion detection under complex conditions.

The efficiency results reveal significant variation among models. UNet, with its lightweight structure, achieved the highest FPS (59.28) but lacked accuracy for complex scenarios. Conversely, SAM2 and Mask2Former demonstrated high segmentation accuracy but relatively low inference speeds (4.09 FPS and 18.46 FPS, respectively), limiting their use in real-time inspection. The proposed method reached 26.56 FPS, offering a favorable balance between accuracy and efficiency, suitable for practical applications requiring both precision and timeliness.

From an application perspective, DeeplabV3+ and UNet are suitable for lightweight deployments, while Segformer and TransUNet provide stronger global semantic modeling for moderately complex inspection tasks. SAM2 and Mask2Former deliver high precision but are better suited for offline analysis or high-performance computing environments. The proposed model, by contrast, achieves high segmentation accuracy with reasonable computational cost, making it well-suited for real-time UAV-based inspection and other edge-deployment scenarios.

In summary, the proposed method consistently outperforms existing mainstream models in segmentation accuracy while maintaining a balanced trade-off between inference speed and computational complexity. Its advantage is particularly evident in corrosion IoU and boundary fitting performance, highlighting its effectiveness for high-precision corrosion segmentation in complex industrial environments.

Additionally, we examined whether the proposed model encountered any typical failure cases across the test set. While the model generally maintained stable performance, minor false positives and missed regions were observed in some challenging scenarios, such as edge-interfered corrosion and fine rust details (see Figure 6). Nevertheless, compared to the baseline model, although there are still minor false positives, the model has shown an improvement in overall detection performance, which demonstrates its robustness and generalization capability.

3.5. Discussion on MCTAM Performance and Scalability

The MCTAM achieves structured fusion of low-level perceptual cues—color and texture—while effectively avoiding feature redundancy or cross-interference. As confirmed by the ablation results, the module contributes to improved segmentation performance under complex background conditions. In addition, due to the parallel design of its sub-branches, MCTAM introduces negligible computational overhead and offers good scalability for practical applications such as real-time UAV-based inspection.

4. Conclusions

This study addresses the challenge of corrosion segmentation in power substation equipment operating under complex environmental conditions and proposes a high-

precision segmentation method based on an improved TransU-Net framework. Specifically, a deformable convolution (DC) module is integrated into the encoder to enhance adaptability to irregular corrosion morphologies; a Multi-Modal Color and Texture Analysis Module (MCTAM) is designed to achieve color–texture feature fusion; and a Shape-IoU loss function is introduced during training to improve boundary fitting accuracy. Together, these components form a robust and generalizable segmentation network.

Experimental results demonstrate that the proposed method outperforms mainstream segmentation models across key evaluation metrics, including mIoU, IoU, and Precision, exhibiting superior performance in complex scenes with blurred corrosion boundaries. Ablation studies confirm the independent contributions and synergistic effects of the proposed modules, validating the effectiveness of multimodal feature fusion and shape-aware optimization. Comparative results further show that the method achieves leading accuracy while maintaining favorable computational efficiency, making it suitable for real-time UAV-based inspection and other practical field applications.

Despite its advantages, several limitations remain. First, the model's performance depends on the diversity and quality of labeled corrosion datasets; when applied to unseen corrosion patterns or lighting conditions, accuracy may degrade due to domain shifts. Second, while the introduction of DC and MCTAM enhances feature representation, it also slightly increases computational complexity, which could challenge deployment on resource-constrained embedded systems. Third, the current Shape-IoU loss function focuses primarily on boundary structure and does not explicitly address inter-class confusion in multi-defect scenarios.

In future work, we plan to address these limitations by exploring lightweight network architectures and domain adaptation strategies to enhance cross-scene generalization. Additionally, integrating the corrosion segmentation outputs with equipment structural features will enable the quantitative assessment of corrosion severity and degradation trends. Ultimately, this line of research aims to establish a closed-loop corrosion monitoring framework that links defect recognition with condition evaluation, thereby providing more reliable support for intelligent operation and refined maintenance of power grid assets.

5. Patents

The technical content related to this article has been filed as a Chinese invention patent entitled "A Method and System for Corrosion Segmentation of Substation Equipment Based on Composite Color and Texture Analysis" (Application No. 2025113997079, filed on 28 September 2025).

Author Contributions: Conceptualization, H.G.; methodology, G.L.; software, H.G. and G.L.; validation, J.G.; formal analysis, J.G.; investigation, Z.L. and X.W.; resources, Z.L. and X.W.; data curation, Z.L. and X.W.; writing—original draft preparation, H.G.; writing—review and editing, H.G. and G.L.; visualization, Z.Z.; supervision, Z.Z.; project administration, Z.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Electric Power Company under Project No. 522702250002, titled "Research and Application on Performance Enhancement of Intelligent Detection for Irregular Visual Defects in Substation Equipment," which was conducted in 2025.

Data Availability Statement: The data presented in this study are available on request from the corresponding author due to confidentiality regulations of the power sector—these regulations are implemented to safeguard the information security of power systems.

Conflicts of Interest: Author Hailong Guo, Guangqi Lu, Xuan Wang are employed by the State Grid Gansu Electric Power Company Tianshui Power Supply Company. Author Jiuyu Guo is employed by the State Grid Gansu Electric Power Company. Author Zhixin Li is employed by the Electric Power Research Institute of State Grid Gansu Electric Power Company. Author Zhenbing Zhao is employed by North China Electric Power University. The authors declare that this study received funding from Electric Power Company. The funder was not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

References

1. Zhang, J.; Zhu, W. Research on Algorithm for Improving Infrared Image Defect Segmentation of Power Equipment. *Electronics* **2023**, *12*, 1588. [CrossRef]
2. Chen, H. Development of a Corrosion Detection and Evaluation System for Typical Substation Equipment. Master's Thesis, Northeast Electric Power University, Jilin, China, 2025.
3. Xu, R.L. Research on Recognition of Rail Fastener Rust Based on Color Image Multi-Threshold Segmentation Algorithm. Master's Thesis, Wuhan University of Technology, Wuhan, China, 2022.
4. Khayatazad, M.; De Pue, L.; De Waele, W. Detection of corrosion on steel structures using automated image processing. *Dev. Built Environ.* **2020**, *3*, 100022. [CrossRef]
5. Vorobel, R.; Ivasenko, I.; Berehulyak, O.; Mandzii, T. Segmentation of rust defects on painted steel surfaces by intelligent image analysis. *Autom. Constr.* **2021**, *123*, 103515. [CrossRef]
6. Shen, H.K.; Chen, P.H.; Chang, L.M. Automated steel bridge coating rust defect recognition method based on color and texture feature. *Autom. Constr.* **2013**, *31*, 338–356. [CrossRef]
7. Zhang, X. Corrosion Defect Detection Algorithm for Power Equipment Based on Deep Learning Object Detection. In Proceedings of the 3rd International Conference on Electrical Power and Energy Systems (ICEPES 2023), Wuhan, China, 15–17 December 2023.
8. Song, W. Corrosion Defect Detection of High-Voltage Transmission Line Vibration Dampers. *J. Electr. Power Sci. Technol.* **2023**, *18*, 78–85.
9. Zhi, Y.L. Recognition of Transmission Line Fitting Corrosion Based on Computer Vision. *Autom. Electr. Power Syst.* **2023**, *47*, 167–174.
10. Chen, F.F.; Dong, H.F.; He, X.Y.; Chen, B.J. Improved U_Net Network for Rust Image Segmentation on Steel Structure Surfaces. *J. Electron. Meas. Instrum.* **2024**, *38*, 49–56.
11. Guo, M.L. Detection Method for Rusty Reinforcement Sections Based on Fully Convolutional Neural Networks. *J. Build. Mater.* **2023**, *26*, 89–94.
12. Chen, F.F.; Cheng, M.T.; Yang, Y.P.; Chen, B.J.; Xiao, W.R. Rust Image Segmentation Fusing Dual Attention Mechanism and U-Net Network. *J. Xi'an Jiaotong Univ.* **2021**, *55*, 119–128.
13. Ju, Z.L. Corrosion Defect Detection for Transmission Lines Based on Multi-Modal Image Fusion. *Power Syst. Technol.* **2023**, *47*, 3056–3063.
14. Zhu, T.; Zhu, S.; Zheng, T.; Ding, H.; Song, W.; Li, C. HEU-Net: Hybrid Attention Residual Block-Based Network with External Skip Connections for Metal Corrosion Semantic Segmentation. *Vis. Comput.* **2024**, *40*, 521–534. [CrossRef]
15. Luo, C.; Yu, L.; Yan, J.; Li, Z.; Ren, P.; Bai, X.; Yang, E.; Liu, Y. Autonomous detection of damage to multiple steel surfaces from 360 panoramas using deep neural networks. *Comput. Aided Civ. Infrastruct. Eng.* **2021**, *36*, 1585–1599. [CrossRef]
16. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. *arXiv* **2021**, arXiv:2102.04306. [CrossRef]
17. Wang, H.X. A Novel Semi-Supervised Extreme Learning Machine and Its Application in Vibration Damper Corrosion Detection. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 1–9. [CrossRef]
18. Zhang, W.; Wang, Y.; Xu, Y.; Zhao, Y.; Liu, Q. FCIHMRT: Feature Cross-Interaction Hybrid Module Residual Network for Corrosion Semantic Segmentation. *Electronics* **2023**, *12*, 4362. [CrossRef]
19. Russell, B.C.; Torralba, A.; Murphy, K.P.; Freeman, W.T. LabelMe: A Database and Web-Based Tool for Image Annotation. *Int. J. Comput. Vis.* **2008**, *77*, 157–173. [CrossRef]
20. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. *Proc. Eur. Conf. Comput. Vis. (ECCV)* **2018**, *11211*, 801–818.
21. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12077–12090.

22. Li, X.; Tian, X.; Wang, Z.; Zhang, F.; Fu, K.; Wang, J. SAM2Former: Segment Anything Model 2 Assisting UNet-Like Transformer for Remote Sensing Image Semantic Segmentation. *IEEE Access* **2025**, *13*, 88765–88777. [CrossRef]
23. Cheng, B.; Misra, I.; Schwing, A.G.; Kirillov, A.; Girdhar, R. Masked-Attention Mask Transformer for Universal Image Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022; pp. 1290–1299.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.