An Improved HRNetV2-Based Semantic Segmentation Algorithm for Pipe Corrosion Detection in Smart City Drainage Networks

Liang Gao *, Xinxin Huang, Wanling Si, Feng Yang *, Xu Qiao, Yaru Zhu, Tingyang Fu, Jianshe Zhao

Posted Date: 18 July 2025

doi: 10.20944/preprints202507.1590.v1

Keywords: semantic segmentation; urban drainage pipeline; HRNetV2; CBAM; pyramid pooling; corrosion detection; smart city; smart infrastructure; deep learning

 doi:10.20944/preprints202507.1590.v1

Article

An Improved HRNetV2-Based Semantic Segmentation Algorithm for Pipe Corrosion Detection in Smart City Drainage Networks

Liang Gao *, Xinxin Huang, Wanling Si, Feng Yang *, Xu Qiao, Yaru Zhu, Tingyang Fu and Jianshe Zhao

School of Artificial Intelligence, China University of Mining and Technology (Beijing), Beijing 100083, China

* Correspondence: gggaoliang1988@163.com (L.G.); yangf@cumtb.edu.cn (F.Y.); Tel.: +86-183-9100-4721 (L.G.)

Abstract

Urban drainage pipelines are essential components of smart city infrastructure, supporting the safe and sustainable operation of underground systems. However, internal corrosion in pipelines poses significant risks to structural stability and public safety. In this study, we propose an enhanced semantic segmentation framework based on High-Resolution Network Version 2 (HRNetV2) to accurately identify corroded regions in Traditional closed-circuit television (CCTV) images. The proposed method integrates a Convolutional Block Attention Module (CBAM) to strengthen the feature representation of corrosion patterns and introduces a Lightweight Pyramid Pooling Module (LitePPM) to improve multi-scale context modeling. By preserving high-resolution details through HRNetV2's parallel architecture, the model achieves precise and robust segmentation performance. Experiments on a real-world corrosion dataset show that our approach attains a mean Intersection over Union (mIoU) of 95.92%, a mean Pixel Accuracy (mPA) of 98.01%, and an overall Accuracy of 98.54%. These results demonstrate the method's effectiveness in supporting intelligent infrastructure inspection and provide technical insights for advancing automated maintenance systems in smart cities.

Keywords: semantic segmentation; urban drainage pipeline; HRNetV2; CBAM; pyramid pooling; corrosion detection; smart city; smart infrastructure; deep learning

1. Introduction

Urban drainage networks form a vital component of smart city infrastructure, re-sponsible for managing stormwater and sewage to sustain urban water cycles and ensure public safety. The structural integrity of these underground pipelines directly impacts ur-ban flood control and environmental health [1]. Due to long-term operation, these pipelines suffer from continuous water flow erosion, internal pressure, corrosive environmental influences [2], and adverse factors such as ground settlement during urban construction, leading to a wide range of structural damages, including corrosion, cracks, and deformation [3]. Among them, internal corrosion is a typical structural defect that gradually erodes the pipe wall material and is a major cause of leakage, rupture, and even collapse. Traditional closed-circuit television (CCTV) inspection relies heavily on manual interpretation of videos by professionals, which suffers from inefficiency, subjectivity, and difficulty in quantitative assessment [4,5]. These limitations hinder the intelligent operation and maintenance of infrastructure required by smart cities. Therefore, the development of high-precision, automated corrosion detection technologies has become a central challenge in smart water management, with significant implications for ensuring underground spatial safety.

Image segmentation, a fundamental task in computer vision, aims to partition images into regions with specific semantics or common features, serving as a foundation for higher-level tasks

 doi:10.20944/preprints202507.1590.v1

such as object recognition and scene understanding. Conventional image segmentation methods based on low-level visual features (e.g., grayscale, color, texture) include thresholding, edge detection, clustering, region growing, and graph-based methods. These approaches have shown effectiveness in constrained scenarios. For example, edge detection methods like the classic Canny operator attempt to delineate object boundaries by identifying abrupt changes in intensity or texture [6]; morphological operations based on pseudo top-hat transforms have been applied to enhance dark-region boundaries and extract pipe defects of specific shapes [7–9]; dynamic thresholding segments images by pixel intensity [10]; graph-based methods model the image as a graph and perform segmentation using graph cuts or partitioning algorithms, applied in water leakage detection [11] and image partitioning [12]; and clustering methods like K-Means group pixels based on similarity in feature space, with applications in corrosion image analysis [13,14]. However, such methods often rely on hand-crafted feature engineering, prior knowledge, and sensitive parameter tuning, making them vulnerable to noise, illumination variation, complex backgrounds, and the diverse morphology of defects. Thresholding is sensitive to uneven lighting; edge detection struggles with noise and incomplete boundaries; region growing depends on seed point selection and growth criteria; and graph-based methods may suffer from high computational complexity and difficult parameterization [12,15].

In recent years, the rapid advancement of deep learning, particularly Convolutional Neural Networks (CNNs), has brought a paradigm shift to the field of image segmentation. Semantic segmentation, which assigns a semantic label to every pixel in an image, has emerged as a foundational technique for pixel-level image understanding. Its ability to precisely delineate object shapes and locations has demonstrated great potential in critical applications such as infrastructure health monitoring and industrial defect inspection [16]. Compared to object detection, semantic segmentation provides pixel-wise granularity, offering a solid technical foundation for subsequent quantitative defect evaluation [17]. Significant progress has been made in applying deep learning–based semantic segmentation to surface defect detection in pipelines and bridge structures. Encoder-decoder architectures, such as U-shaped convolutional networks (U-Net) and its variants, have become mainstream solutions due to their excellent feature fusion capabilities. U-Net, proposed by Ronneberger et al. [18], introduced skip connections to effectively fuse high- and low-level features and has since become a leading framework for pipe defect detection. Subsequent works have further improved performance by integrating attention mechanisms such as the Convolutional Block Attention Module (CBAM), Coordinate Attention (CA), and Squeeze-and-Excitation (SE), enabling networks to focus more effectively on defect regions while suppressing interference from complex pipeline backgrounds [19,20]. Models such as DeepLabv3+, SparseInst, and instance segmentation architectures have been employed to achieve more refined localization and differentiation of defects [21]. Liu et al. [22] enhanced DeepLabv3+ using Atrous Spatial Pyramid Pooling (ASPP) to improve multi-scale defect recognition, incorporating Efficient Channel Attention (ECA) modules in the encoder-decoder to focus on critical information, thereby boosting feature representation and segmentation accuracy. Wang et al. [23] designed the SparseInst framework for high-precision instance-level defect segmentation, introducing TensorRT acceleration to meet real-time inspection demands. Moreover, ensemble learning strategies, data augmentation via Generative Adversarial Networks, multimodal fusion, and Conditional Random Field (CRF)–based post-processing have been extensively explored to improve robustness, generalization, and boundary precision. Forkan et al. [24] proposed CorrDetector, employing ensemble learning for Unmanned Aerial Vehicle (UAV) corrosion detection with strong resilience to noise and large-scale variations. Li et al. [25] leveraged Style-Based Generative Adversarial Network 3 (StyleGAN3) to synthesize corrosion images for addressing data scarcity and designed an enhanced DeepLabv3+ to achieve improved segmentation under few-shot conditions. Papamarkou et al. [26] applied residual networks to automatically identify corrosion in dry nuclear fuel storage casks, combining preprocessing and residual connections to boost detection accuracy with minimal manual annotation. Jin et al. [27] integrated sonar and optical imaging for multimodal underwater defect recognition and guided the network

 doi:10.20944/preprints202507.1590.v1

using channel and spatial attention to enhance multi-defect identification, particularly in blurry and low-contrast sonar images. Wang [28] incorporated CRF to refine segmentation boundaries, improving contour accuracy and consistency, especially for corrosion-crack hybrid defect scenarios.

The core advantage of deep learning lies in its end-to-end feature learning capability, enabling it to learn highly discriminative hierarchical features directly from raw data, thus overcoming the limitations of manually crafted features in traditional methods [29,30]. Numerous studies have confirmed that deep learning models significantly outperform traditional approaches in terms of segmentation accuracy, robustness, and generalization across challenging scenarios involving illumination variations, noise, cluttered backgrounds, and morphological diversity of pipeline defects. These models have been successfully applied in UAV inspection, industrial endoscopy, CCTV-based sewer inspection [31], and nuclear facility monitoring. For instance, Nash et al. [32] constructed a corrosion annotation dataset through crowdsourcing and used CNNs for automated detection of multiple corrosion types in industrial settings, reducing annotation cost and enhancing model generalization. Subsequently, Burton et al. [33] proposed the RustSEG semantic segmentation framework, combining U-Net and Residual Network (ResNet) encoders for pixel-wise corrosion segmentation. Their model achieved up to 85% mean Intersection over Union (mIoU) even under oil contamination and severe rust interference, effectively addressing challenges such as fuzzy boundaries and small defect detection, demonstrating strong practical value. Despite these advances, deploying deep-learning-based semantic segmentation in real-world urban drainage pipeline inspection remains challenging. Firstly, obtaining a large number of high-quality, pixel-level labeled defect images is costly and labor-intensive, limiting model training. Secondly, complex background elements—such as pipe textures, water residue, sludge deposits, and structural shadows—often resemble real defects, leading to false positives and missed detections, requiring models to have stronger contextual reasoning and anti-interference capabilities. Thirdly, fine-grained defects (e.g., cracks, pitting, patchy corrosion) exhibit diverse scales and shapes, demanding models with effective multi-scale feature extraction capabilities. However, traditional U-Net models suffer from high-frequency detail loss during downsampling, resulting in blurred boundaries. DeepLabv3+–like networks often destroy local textures due to dilated convolutions. Furthermore, segmentation performance may vary under different pipe materials, aging conditions, imaging devices, or environmental settings, revealing limitations in model robustness and transferability.

To address the above challenges, especially the difficulty of detecting small-scale corrosion in complex backgrounds and the risk of background-induced false detections, this study proposes an improved semantic segmentation framework based on the High-Resolution Network Version 2 (HRNetV2). The main contributions of this paper are summarized as follows:

1. An improved HRNetV2 architecture tailored for precise segmentation of corrosion regions in complex drainage pipe environments is proposed. For the first time, this framework systematically integrates the CBAM and a Lightweight Pyramid Pooling Module (LitePPM), jointly improving the network's ability to perceive multi-scale defects and resist background interference.
2. Inspired by the Pyramid Scene Parsing Network (PSPNet), a LitePPM is designed. This module uses adaptive average pooling at multiple scales, followed by convolution and feature concatenation to extract and fuse contextual information, expanding the network's receptive field while controlling parameter growth.
3. Comprehensive experiments on a self-built drainage pipe defect dataset validate the effectiveness of the proposed approach. Results show that the model significantly outperforms mainstream U-Net variants and the original HRNetV2 in terms of segmentation accuracy, mIoU and recall.

 doi:10.20944/preprints202507.1590.v1

2. Materials and Methods

2.1. Drainage Pipeline Data Acquisition

The dataset used in this study originates from the corrosion images of drainage pipelines captured by a CCTV pipe inspection robot developed at China University of Mining and Technology (Beijing). The original images were screened by professionals to ensure quality. These images cover a diverse range of viewpoints, including front views, side views, and partial views of the pipeline, totaling 1,360 corrosion images. This dataset is highly suitable for corrosion area segmentation tasks in drainage pipelines. Pixel-level annotations were carried out using the LabelMe software, where the corrosion regions were labeled as "FS." The annotated data was then converted into the VOC dataset format. Finally, the dataset was divided into training, validation, and testing sets at a ratio of 8:1:1. Figure 1(a) presents examples of the original images showing corrosion defects in drainage pipelines, and Figure 1(b) shows the corresponding annotated label images, where black represents the background and white indicates the corroded regions.

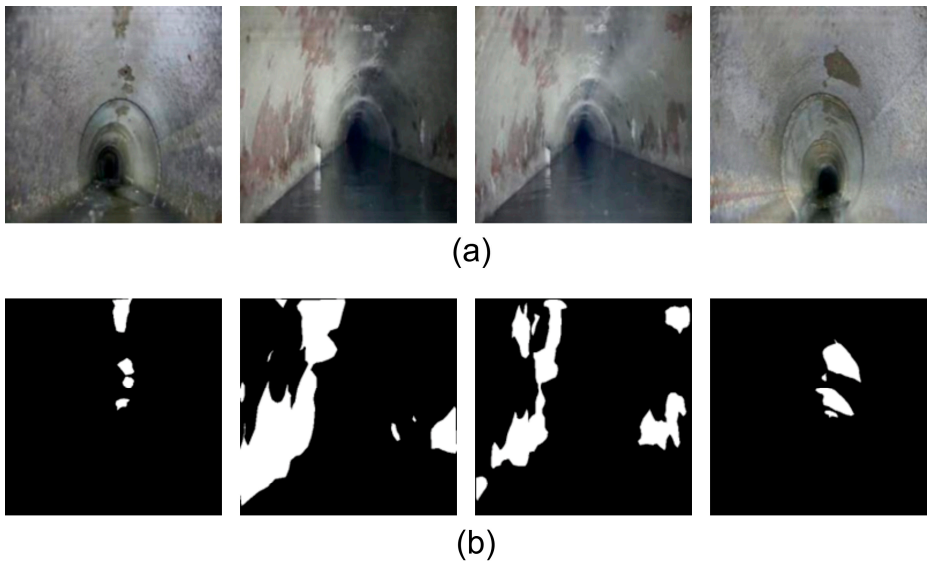

Figure 1. Original and annotated images of partial pipeline corrosion: **(a)** Original images of corrosion defects in drainage pipelines; **(b)** Annotated label images.

2.2. Data Augmentation

To improve the model's generalization ability in the corrosion area segmentation task for drainage pipelines, a variety of data augmentation strategies were employed during the data loading stage. These include random scaling and cropping, horizontal flipping, Gaussian blurring, random rotation, and color distortion. These augmentation techniques are applied in random combinations to each pair of input image and label, which effectively increases the diversity of training samples and reduces the risk of model overfitting. These techniques are highly controllable, flexible, and computationally efficient with low implementation complexity. Examples of the augmented corrosion area images and their corresponding labels are shown in Figure 2.

 doi:10.20944/preprints202507.1590.v1

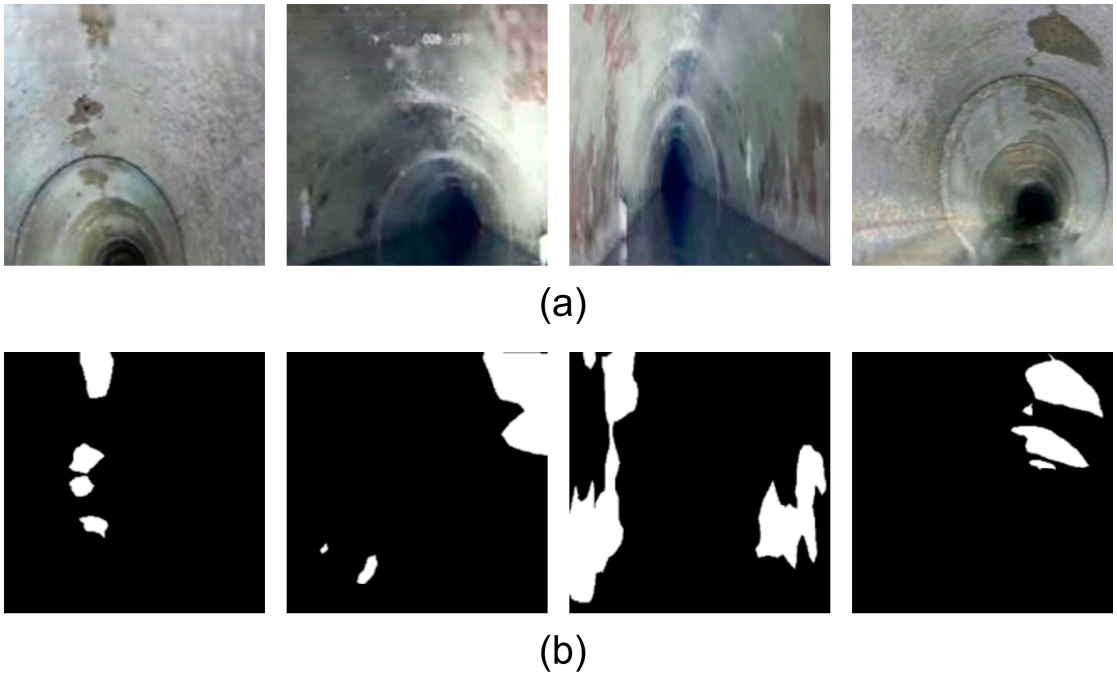

Figure 2. Enhanced original and annotated images of partial pipeline corrosion: **(a)** Enhanced original images; **(b)** Enhanced annotated label images.

2.3. Semantic Segmentation Network for Corrosion Areas in Drainage Pipelines

To address the challenges encountered by traditional networks in such segmentation tasks, this study proposes a multi-branch high-resolution parallel network based on the HRNetV2 architecture. The framework of the proposed method is shown in Figure 3. First, raw image data of internal sewer defects are acquired using the CCTV inspection robot. The collected images are then preprocessed by resizing them and padding gray borders to match the network input dimensions. Based on this, data augmentation techniques such as random scaling, horizontal flipping, translation cropping, Gaussian blurring, random rotation, and color disturbance are applied to further enrich the dataset and enhance the model's generalization. The dataset is divided into training, validation, and testing sets in an 8:1:1 ratio. This study modifies the baseline HRNetV2 by integrating the CBAM attention mechanism to enhance the feature response of key areas and embeds the LitePPM module during high-level semantic feature extraction to achieve multi-scale information fusion and aggregation. Finally, the predicted segmentation maps output by the model enable pixel-level identification of corroded regions within the sewer. Model performance is validated and analyzed using various evaluation metrics on the test set.

 doi:10.20944/preprints202507.1590.v1

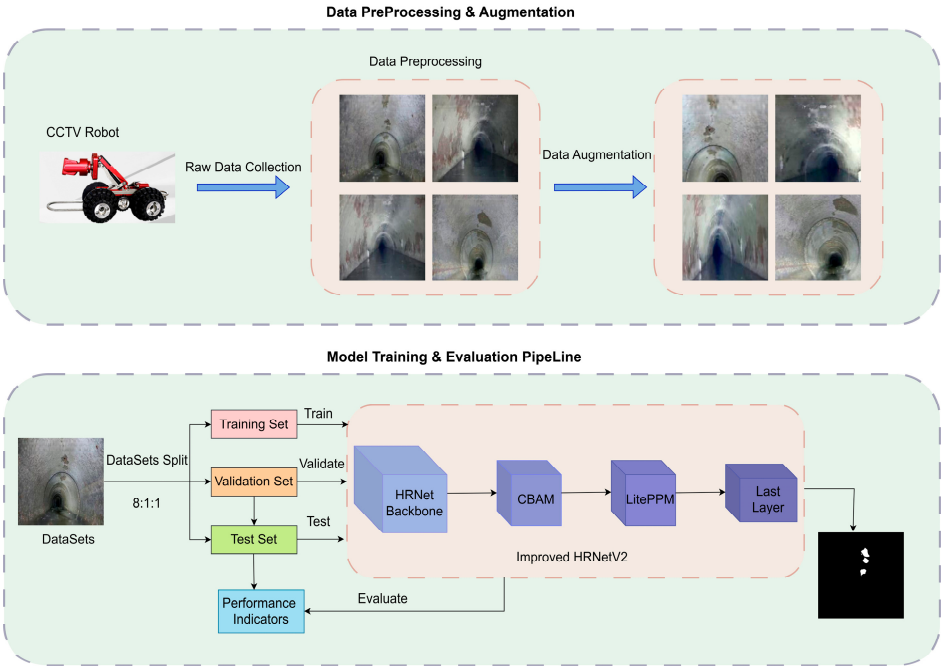

Figure 3. Workflow of the proposed method.

2.3.1. HRNetV2 Semantic Segmentation Model

High-Resolution Network (HRNet) was first proposed by Sun et al. [34] at CVPR 2019. Its core innovation lies in breaking away from the traditional encoder-decoder structure by constructing parallel multi-resolution subnetworks that maintain high-resolution representations throughout the network. The HRNet series includes High-Resolution Network Version 1 (HRNetV1), HRNetV2, High-Resolution Network Version 2 Plus (HRNetV2P). HRNetV1 outputs only high-resolution branch features, lacking deep semantic utilization. Later that year, Sun et al. proposed HRNetV2 and HRNetV2P for segmentation and detection tasks, respectively [35]. HRNetV2 fuses deep semantic features from all branches while preserving details and providing global semantic context information.

The network structure of the HRNetV2 semantic segmentation model is illustrated in Figure 4. It consists of three parts: backbone feature extraction, feature integration, and prediction output. The backbone includes four stages. The input image size is 480×480×3. In the first stage, two 3×3 convolutions with a stride of 2 downsample the image to 120×120 and increase the channel number to 64. This is followed by four standard ResNet bottleneck blocks, each comprising two 1×1 convolutions and one 3×3 convolution, increasing the channel number to 256. The resulting feature map size is 120×120×256. The second stage introduces two parallel branches with resolutions of 120×120 and 60×60 and channel numbers of 32 and 64, respectively. Each branch contains four BasicBlocks (each with two 3×3 convolutions) for scale-specific feature extraction. Cross-branch fusion modules share multi-scale features. The resulting feature maps are 120×120×32 and 60×60×64. In stages 3 and 4, the number of branches expands to three and four, respectively. Each branch continues with four BasicBlocks, and multi-scale fusion is performed at the end of each stage. The feature maps for stage 3 are 120×120×32, 60×60×64, and 30×30×128. For stage 4, they are 120×120×32, 60×60×64, 30×30×128, and 15×15×256.

 doi:10.20944/preprints202507.1590.v1

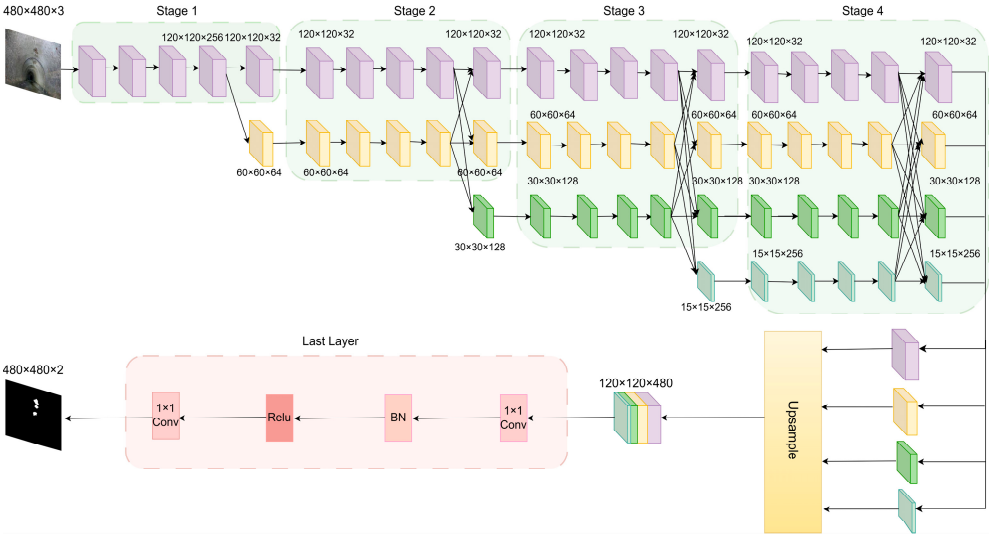

Figure 4. Network architecture of the HRNetV2 model.

2.3.2. Optimization of HRNetV2 Semantic Segmentation Model

To further enhance the performance of HRNetV2 in the task of corrosion region segmentation in drainage pipelines, this study introduces structural improvements targeting its limitations in complex environments. Although HRNetV2 possesses a multi-branch parallel structure that maintains high-resolution feature representation, its original design lacks an explicit attention mechanism. As a result, the model struggles to effectively distinguish critical features from redundant background information—particularly under challenges such as blurred corrosion boundaries, unclear textures, and severe background interference. In this study, CBAM modules are incorporated into the outputs of the four branches in the stage 4 of the HRNetV2 backbone. The rationale is as follows: First, stage 4 is the final multi-scale feature extraction stage in HRNetV2. Its four branch outputs correspond to different spatial resolutions (120×120×32, 60×60×64, 30×30×128, and 15×15×256), and the features have already integrated multi-level semantic information. By adding CBAM to stage 4, the dual attention mechanism—channel and spatial—can be fully leveraged to enhance the response of key regions, ensuring that corrosion-related features are emphasized at each scale while suppressing background noise. Second, the CBAM module is lightweight and does not introduce significant parameter overhead, making it suitable for feature enhancement after parallel branches. Moreover, considering that the original HRNetV2 lacks a global contextual modeling mechanism after multi-scale feature fusion—leading to misclassification or omission in the presence of corrosion regions with varying scales and complex shapes—this study introduces a custom-designed LitePPM module after upsampling and channel-wise concatenation of all feature branches. The reasons for placing LitePPM after the fusion step are twofold: First, The fused feature map integrates spatial details and semantic information from multiple scales. Inserting LitePPM at this point enables effective capture of global semantic dependencies, thereby enhancing the model's ability to perceive large-scale or spatially varying corrosion regions and significantly improving the recall rate. Second, Compared to embedding a context module directly in the backbone, inserting LitePPM here serves a clearer global modeling purpose, contributing to the model's improved recognition of corrosion areas in complex backgrounds. Finally, the context-enhanced fused features are passed through a 1×1 convolution layer for channel compression and then fed into a Softmax classifier to generate binary segmentation results, achieving pixel-wise prediction of the corrosion regions. The improved network architecture is illustrated in Figure 5.

 doi:10.20944/preprints202507.1590.v1

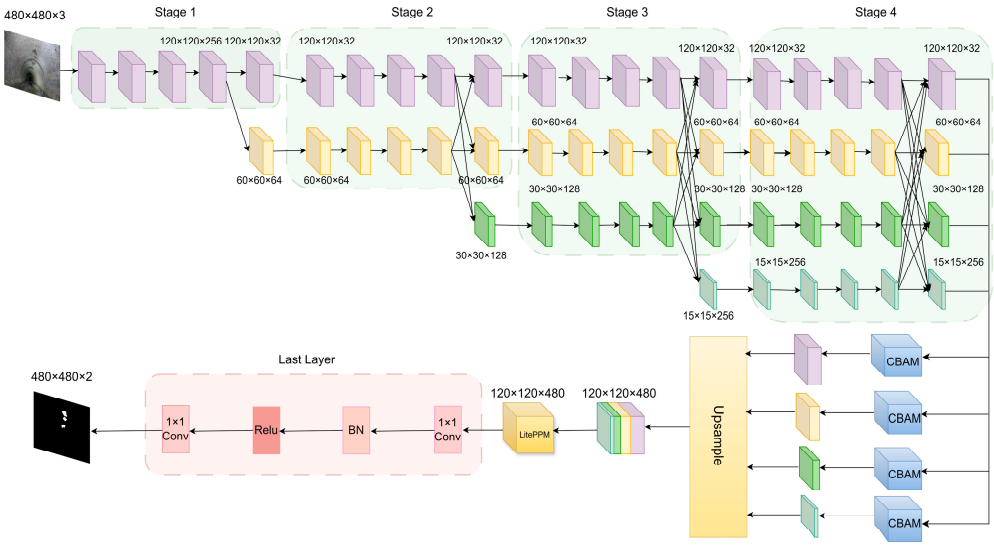

Figure 5. Improved HRNetV2 model .

2.3.3. Integration of the CBAM Attention Mechanism

To improve the segmentation accuracy of corroded regions in drainage pipelines, this study integrates the Convolutional Block Attention Module (CBAM). This mechanism aims to address the challenge of weak responses of corrosion features within complex backgrounds, where critical information is often overwhelmed by irrelevant features. CBAM was first introduced by Woo et al. [36] in 2018, and its core idea is to guide the network's attention towards more informative features by explicitly modeling both channel and spatial attention. As shown in Figure 6, the CBAM module consists of two sequential submodules: the Channel Attention Module (CAM) and the Spatial Attention Module (SAM). The CAM is designed to capture the importance of each feature channel by leveraging both global average pooling and global max pooling, followed by a shared Multi-Layer Perceptron (MLP). The spatial attention module, in contrast, focuses on "where" to emphasize by enhancing the network's sensitivity to critical spatial regions.

The CAM focuses on mining the response differences among channels to generate a channel-wise importance weight map, thereby enhancing the network's ability to model salient semantic features. Specifically, global average pooling and global max pooling are first applied to the input feature map $F \in R^{C\times H\times W}$, resulting in two global context descriptors, z_{avg} and z_{max}. These descriptors are then passed through a shared MLP composed of two fully connected layers with intermediate dimensionality reduction and expansion controlled by a compression ratio r. The outputs are combined and activated using a Sigmoid function to obtain the channel attention weight map M_c. This weight map is multiplied with the input feature map on a channel-wise basis to perform channel attention recalibration. The computation is expressed as:

$$M_c = \sigma\left(W_2 \times \delta\left(W_1 \times z_{avg} + W_1 \times z_{max}\right)\right) \quad (1)$$

where $z_{avg=GAP(F)\in R^{C\times 1\times 1}}$, $z_{max=GMP(F)\in R^{C\times 1\times 1}}$, $W_1 = R^{\frac{C}{r}\times C}, W_2 = R^{C\times\frac{C}{r}}$, δ denotes ReLU activation function, σ denotes the Sigmoid function.

The SAM concentrates on "which spatial regions" deserve more attention, thereby improving the model's ability to discriminate local areas. Taking as input the feature map refined by the channel attention, it first performs average pooling and max pooling operations along the channel dimension, generating two single-channel spatial descriptors. These are concatenated along the channel axis and passed through a convolutional layer with a 7×7 kernel, followed by a Sigmoid function to generate the spatial attention map M_c. Finally, the spatial attention map is multiplied element-wise with the input feature map to enhance spatially relevant locations. The calculation is as follows:

 doi:10.20944/preprints202507.1590.v1

$$M_s = \sigma\left(f^{7\times7}([AP(F);\ MP(F)])\right) \tag{2}$$

where $f^{7\times7}$ represents the convolution operation with a 7×7 kernel, AP and MP denote the average pooling and max pooling along the channel dimension, respectively, and $M_s \in R^{1\times H\times W}$ denotes the Sigmoid function.

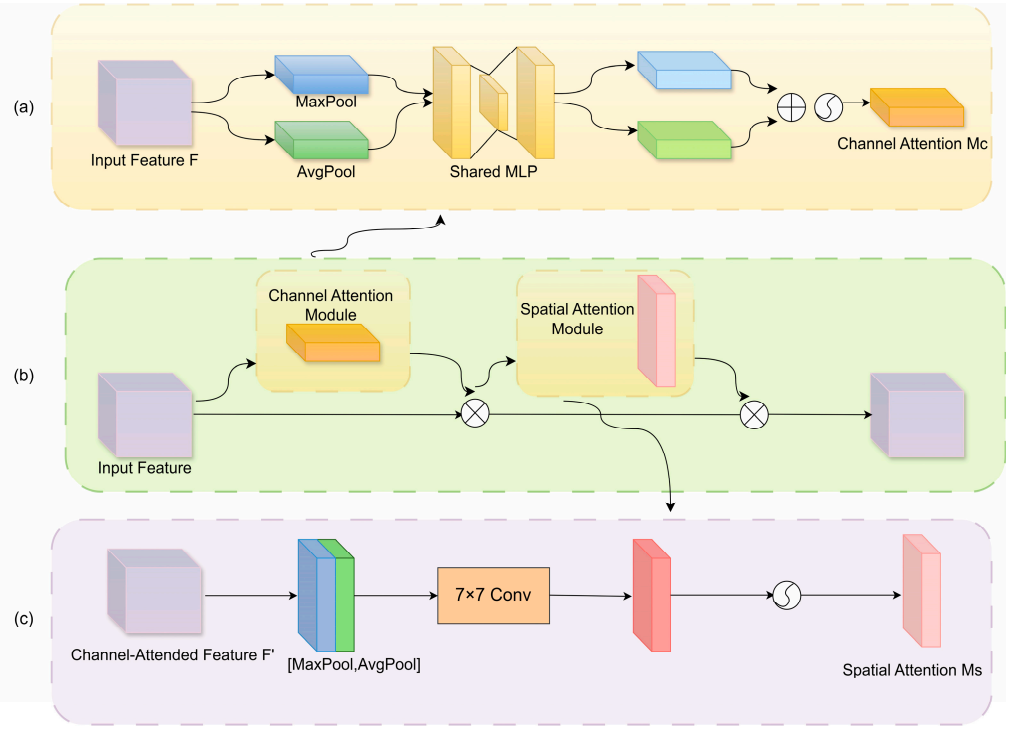

Figure 6. Structure of CBAM: **(a)** Channel attention module; **(b)** CBAM; **(c)** Spatial attention module.

2.3.4. Introduction of the LitePPM Module

In complex backgrounds with small targets, traditional segmentation networks often struggle due to limited receptive fields or loss of semantic information at high levels. In real-world pipeline inspections, corrosion areas vary greatly in shape and size, requiring strong contextual awareness. To address this, the LitePPM module is introduced after multi-scale feature fusion. Inspired by the pyramid pooling module in PSPNet [37], LitePPM adopts a lightweight design to provide effective multi-scale context extraction with low parameter overhead. The module includes a multi-scale adaptive pooling stage and an upsampling and fusion stage, as shown in Figure 7.

In the first stage, adaptive average pooling is applied at multiple scales (1×1, 2×2, 3×3, 6×6) to extract contextual information with varying receptive fields. Each pooled feature is compressed via a 1×1 convolution, normalized using BatchNorm, and activated with ReLU. In the second stage, all features are upsampled using bilinear interpolation to the original size and concatenated with the original feature map. Finally, a 3×3 convolution block fuses the concatenated features to produce the enhanced context feature map.

 doi:10.20944/preprints202507.1590.v1

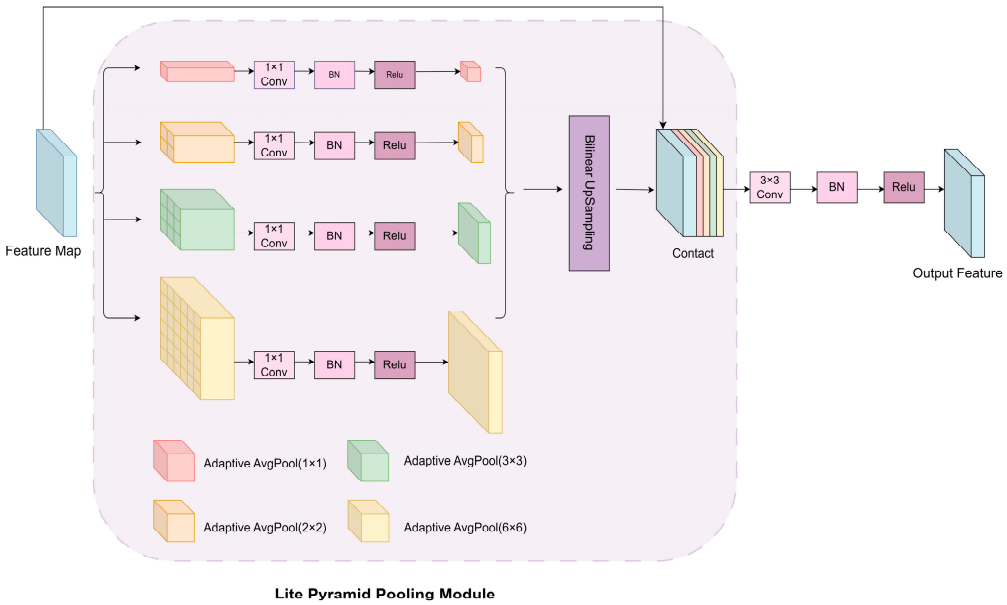

Figure 7. Structure of the LitePPM module.

3. Experiments and Analysis

3.1. Experimental Platform and Parameters

All experiments in this study were conducted under the Windows 10 environment. The hardware configuration includes an Intel(R) Xeon(R) Gold 6133 CPU @ 2.50GHz processor, 256GB of RAM, and an NVIDIA GeForce RTX 4090 GPU with 24GB of memory. CUDA 12.6 was used for parallel acceleration, and the programming language was Python 3.8 with the PyTorch 2.2.0 deep learning framework. During model training, hrnetv2_w32 was selected as the backbone feature extraction network, and the number of output channels was set to 2 for binary segmentation of corrosion and background regions. Hyperparameter tuning was conducted to optimize model performance, with the final settings listed in Table 1.

Table 1. Tuned hyperparameters.

Hyperparameter	Value
Epoch	221
Batch Size	5
Optimizer	SGD
Momentum	0.9
Init Learning Rate	4e-3
Min Learning Rate	4e-5
Learning Rate Decay Type	COS
Weight Decay	1e-4

3.2. Loss Function

To improve performance in segmenting corrosion areas under sample imbalance and enhance boundary precision, a combined loss function is designed. It consists of Dice Similarity Coefficient Loss (Dice Loss) and Focal Loss, equally weighted.

 doi:10.20944/preprints202507.1590.v1

Dice Loss is derived from the Sørensen–Dice coefficient [38] and is commonly used in medical imaging and semantic segmentation. It optimizes overlap similarity and alleviates training bias caused by class imbalance. The Dice Loss is defined in Equation (3):

$$Dice\ Loss = 1 - \frac{2\sum_{i=1}^{N} p_i g_i + \varepsilon}{\sum_{i=1}^{N} p_i + \sum_{i=1}^{N} g_i + \varepsilon} \tag{3}$$

where p_i is the predicted pixel value, g_i is the ground truth label, N is the number of pixels, and ε is a small constant to prevent division by zero. Dice Loss maximizes the intersection between prediction and ground truth, improving sensitivity to edges and small regions.

Focal Loss, proposed by Lin et al. [39], addresses extreme class imbalance by down-weighting well-classified examples and focusing on hard examples. It is particularly suitable for cases where the target occupies a small area. Equation (4) presents the definition of the Focal Loss:

$$Focal\ Loss = -\alpha(1 - p_t)^{\gamma} \log(p_t) \tag{4}$$

where p_t is the predicted probability of the true class, $\alpha \in (0,1)$ balances class weights, and $\gamma \geq 0$ is the focusing parameter. Typically, $\gamma = 2, \alpha = 0.25$. In this study, γ is set to its default value of 0, class weights are set as 1:1 for foreground and background.

The final loss function is defined in Equation (5):

$$Total\ Loss = \lambda_1 Dice\ Loss + \lambda_2 Focal\ Loss \tag{5}$$

where $\lambda_1 = \lambda_2 = 1$, meaning both losses contribute equally. This combination accelerates convergence and improves boundary discrimination, providing better robustness in detecting blurred edges and low-contrast corrosion regions.

3.3. Evaluation Metrics

To comprehensively evaluate the model's performance on the drainage pipeline corrosion segmentation task, several mainstream semantic segmentation metrics were employed, including mPrecision, Accuracy, mIoU, mean Pixel Accuracy (mPA).

$$mPrecision = \frac{1}{n}\sum_{i=1}^{n} \frac{TP_i}{TP_i + FP_i} \tag{6}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

$$mIoU = \frac{1}{n}\sum_{i=1}^{n} \frac{TP_i}{TP_i + FP_i + FN_i} \tag{8}$$

$$mPA = \frac{1}{n}\sum_{i=1}^{n} \frac{TP_i}{TP_i + FN_i} \tag{9}$$

where TP denotes the number of pixels correctly predicted as corrosion, FP denotes background pixels incorrectly predicted as corrosion, FN represents corrosion pixels incorrectly predicted as background, and TN represents background pixels correctly predicted as background.

3.4. Experimental Results and Analysis

To verify the effectiveness of the proposed improved model in semantic segmentation tasks, we compared its performance with several mainstream models, including DeepLabV3+, PSPNet, and U-Net. All models were trained and evaluated under the same dataset and training strategies. The results are presented in Table 2.

PSPNet and DeepLabV3+ demonstrated excellent performance supported by the lightweight backbone MobileNetV2, achieving mIoU values of 94.62% and 95.24%, and mPA values of 97.48% and 97.84%, respectively. However, these models exhibited limitations in handling fine structures

 doi:10.20944/preprints202507.1590.v1

and boundary regions, with occasional misclassification. U-Net, when combined with ResNet50 and VGG16 as backbones, further enhanced feature extraction capabilities and demonstrated strong accuracy and stability. ResNet50 was preferred as the backbone due to its deep residual structure, which effectively mitigates gradient vanishing and enhances multi-scale semantic expression. When using ResNet50, U-Net achieved mIoU, mPA, Accuracy, and mPrecision of 95.79%, 97.94%, 98.49%, and 97.73%, respectively. Compared to DeepLabV3+, mIoU improved by 0.58%, indicating better pixel-level classification accuracy within the target regions. HRNetV2, with its high-resolution feature preservation and multi-scale fusion mechanism, achieved superior spatial localization and global contextual awareness. Among HRNetV2 variants, HRNetV2_W32 outperformed HRNetV2_W18 due to its larger number of channels and stronger representational power, with mIoU and Accuracy reaching 95.82% and 98.51%. After integrating attention mechanisms and context enhancement modules, the improved HRNetV2_W32 model achieved the highest performance: mIoU of 95.92%, mPA of 98.01%, Accuracy of 98.54%, and mPrecision of 97.80%, confirming the effectiveness and adaptability of the proposed method.

Table 2. Evaluation metrics of different models.

Network	Backbone	mIoU	mPA	Accuracy	mPrecision
DeepLabv3+	MobileNetV2	95.24	97.84	98.29	97.27
PSPNet	MobileNetV2	94.62	97.48	98.05	96.96
U-Net	ResNet50	95.79	97.94	98.49	97.73
U-Net	VGG16	95.73	97.82	98.48	97.80
HRNetV2	HRNetV2_W18	95.56	97.80	98.41	97.63
HRNetV2	HRNetV2_W32	95.82	97.92	98.51	97.78
Improved HRNetV2	**HRNetV2_W32**	**95.92**	**98.01**	**98.54**	**97.80**

The results of the improved method are shown in bold.

Figure 8 illustrates the segmentation results of different models. The first row displays the original images, the second row shows the manually annotated ground truth labels, and the third to seventh rows present the segmentation outputs of PSPNet, DeepLabV3+, U-Net, the original HRNetV2, and the proposed improved HRNetV2 model, respectively. From the results, PSPNet demonstrates certain robustness in detecting large-scale corrosion regions, successfully outlining the main contours of the corrosion areas. However, it exhibits significant shortcomings in boundary handling and small-object detection. In Figure 8 (c), some fine corrosion areas are missed, while Figure 8 (b) shows adhesion issues in adjacent regions. DeepLabV3+ outperforms PSPNet in terms of regional integrity and can more effectively identify medium-sized corrosion areas. However, it still tends to miss or falsely detect regions with blurry edges or small targets. For instance, in Figure 8 (e), the boundaries of corrosion regions are not clearly defined, and in Figure 8 (c), the small corrosion regions are not segmented at all. U-Net, with its encoder-decoder architecture, exhibits good continuity and structural recovery in medium-sized corrosion area detection. Nonetheless, its performance in recognizing small corrosion areas, such as in Figure 8 (c), is suboptimal. It tends to over-smooth features, which negatively impacts boundary precision. In Figure 8 (a), it also fails to segment corrosion regions with indistinct edges. The original HRNetV2 model, utilizing a parallel multi-branch structure, shows certain advantages in preserving segmentation details and maintaining high-resolution feature extraction. In Figures 8 (a) and (c), its output offers clearer edge retention compared to the models mentioned above and demonstrates better detail recovery than U-Net and DeepLabV3+. However, due to the lack of explicit attention mechanisms and contextual information enhancement modules, it still struggles in distinguishing regions under complex texture interference, with some minor corrosion areas—such as those in Figure 8(e)—being missed. In contrast, the proposed improved HRNetV2 model achieves the best segmentation results across all samples. Benefiting from the CBAM module, which enhances the perception of key regions, and the LitePPM module, which improves multi-scale contextual modeling, the model can more accurately

 doi:10.20944/preprints202507.1590.v1

capture the edges and fine details of corrosion areas. The predictions in the seventh row are highly consistent with the ground truth labels, with the shape, location, and boundaries of corrosion regions being more precisely represented. The model significantly outperforms other comparative methods in both segmentation accuracy and stability.

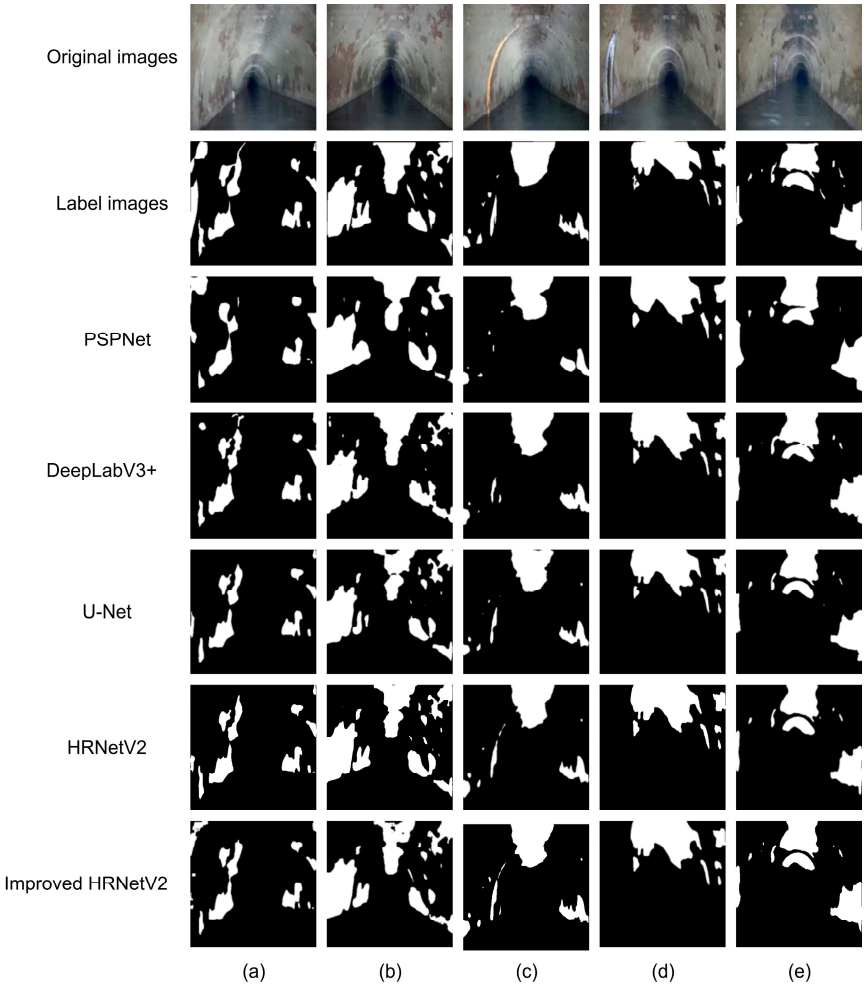

Figure 8. Segmentation results of different models.

3.5. Ablation Study and Grad-CAM Visualization Analysis

To further investigate the contribution and focus of different modules in the improved model, ablation studies and Grad-CAM visualizations were conducted to assess feature responsiveness both qualitatively and quantitatively.

In terms of quantitative evaluation, Table 3 shows that the baseline HRNetV2 model achieved mIoU of 95.82%, mPA of 97.92%, Accuracy of 98.51%, and mPrecision of 97.78%. After adding the CBAM module, the model's ability to focus on critical areas improved, slightly increasing all metrics. Introducing the LitePPM module alone enhanced context modeling, leading to an mIoU of 95.98% and mPA of 98.11%. The combined model with both CBAM and LitePPM achieved mIoU of 95.92% and mPA of 98.01%, showing that the two modules are complementary and improve overall performance.

 doi:10.20944/preprints202507.1590.v1

Table 3. Ablation study results.

HRNetV2	CBAM	LitePPM	mIoU	mPA	Accuracy	mPrecision
√	×	×	95.82	97.92	98.51	97.78
√	√	×	95.85	97.94	98.52	97.79
√	×	√	95.98	98.11	98.53	97.83
√	√	√	95.92	98.01	98.54	97.80

For qualitative analysis, Figure 9 presents Grad-CAM visualizations: the first row shows baseline HRNetV2 with scattered and less accurate focus; the second row (with CBAM) shows improved attention to corrosion edges; the third row (with LitePPM) highlights stronger, more continuous focus on corrosion regions, proving its effectiveness in context modeling and localization.

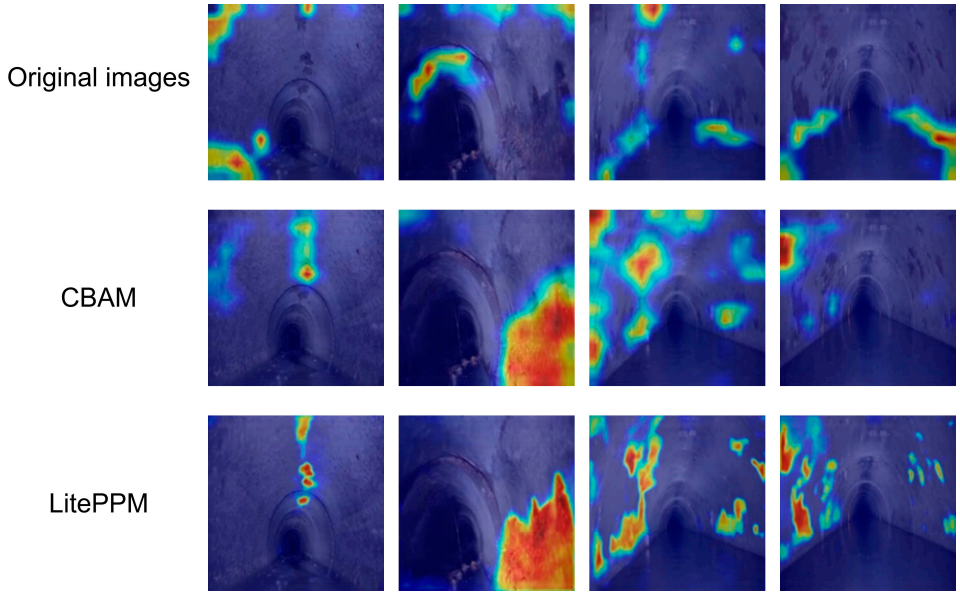

Figure 9. Grad-CAM visualization of attention across different architectures.

4. Discussion

This study proposes an improved HRNetV2-based network to address challenges in segmenting corroded regions in drainage pipelines, such as low precision, missed or false detections under complex backgrounds, and limited contextual modeling. Random data augmentation, including cropping, rotation, and color jittering, was applied to enhance robustness and generalization. CBAM and LitePPM modules were introduced to strengthen feature extraction and context modeling. Extensive experiments on a self-constructed dataset demonstrated that the improved model outperforms the original HRNetV2 and classic models like DeepLabV3+, PSPNet, and U-Net in mIoU, mPA, Accuracy, and mPrecision. The improved model achieved an mIoU of 95.92%, surpassing PSPNet, DeepLabV3+, U-Net, and baseline HRNetV2 by 1.37%, 0.71%, 0.14%, and 0.10%, respectively. It exhibited excellent performance in identifying corrosion regions and preserving edge details. Furthermore, Grad-CAM visualization confirmed that CBAM and LitePPM modules significantly enhance the model's ability to focus on critical features and express contextual information, improving interpretability and reliability of the network. These advancements contribute valuable tools for the intelligent inspection and maintenance of urban infrastructure, supporting the development of smart city management systems aimed at ensuring sustainable and safe urban environments.

Author Contributions: Conceptualization, L.G.; methodology, L.G. and X.H.; validation, L.G., X.H. and W.S.; formal analysis, L.G. and X.H.; investigation, L.G. and X.Q.; resources, F.Y. and X.Q.; funding acquisition, F.Y.;

 doi:10.20944/preprints202507.1590.v1

data curation, W.S., Y.Z. and L.G.; writing—original draft preparation, L.G.; writing—review and editing, L.G. and X.H.; visualization, L.G.; supervision, T.F. and Y.Z.; project administration, J.Z. and W.S.; All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the National Natural Science Foundation of China under Grant No. 52427901. This work also received support from the Fundamental Research Funds for the Central Universities (Ph.D. Top Innovative Talents Fund of CUMTB) under Grant No. BBJ2025073.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset used in this study originates from China University of Mining and Technology (Beijing), and the authors have obtained permission to use the data. For access to the dataset, please contact the corresponding author directly.

Acknowledgments: The authors sincerely thank the team members who contributed to image acquisition, annotation, and experimental support. Their assistance was essential to the successful completion of this work. The authors also appreciate the valuable comments and constructive suggestions provided by the anonymous reviewers, which greatly improved the quality of the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

HRNetV2	High-Resolution Network Version 2
CBAM	Convolutional Block Attention Module
LitePPM	Lightweight Pyramid Pooling Module
mIoU	mean Intersection over Union
mPA	mean Pixel Accuracy
CCTV	Traditional closed-circuit television
CNNs	Convolutional Neural Networks
U-Net	U-shaped convolutional networks
CA	Coordinate Attention
SE	Squeeze-and-Excitation
ASPP	Atrous Spatial Pyramid Pooling
ECA	Efficient Channel Attention
CRF	Conditional Random Field
UAV	Unmanned Aerial Vehicle
StyleGAN3	Style-Based Generative Adversarial Network 3
ResNet	Residual Network
PSPNet	Pyramid Scene Parsing Network
HRNet	High-Resolution Network
HRNetV1	High-Resolution Network Version 1
HRNetV2P	High-Resolution Network Version 2 Plus
SAM	Spatial Attention Module
CAM	Channel Attention Module
MLP	Multi-Layer Perceptron
Dice Loss	Dice Similarity Coefficient Loss

References

1. Shen, D.; Liu, X.; Shang, Y.; Tang, X. Deep Learning-Based Automatic Defect Detection Method for Sewer Pipelines. *Sustainability* **2023**, *15*, 9164, doi:10.3390/su15129164.
2. Yuan, G.; Hong-Wu, W.; Shan-Fa, Z.; Lu-Ming, M.A. Current Research Progress in Combined Sewer Sediments and Their Models. *China Water & Wastewater* **2010**.

 doi:10.20944/preprints202507.1590.v1

3. Yin, X.; Chen, Y.; Bouferguene, A.; Zaman, H.; Al-Hussein, M.; Kurach, L. A Deep Learning-Based Framework for an Automated Defect Detection System for Sewer Pipes. *Automation in construction* **2020**, *109*, 102967.
4. Cheng, J.C.; Wang, M. Automated Detection of Sewer Pipe Defects in Closed-Circuit Television Images Using Deep Learning Techniques. *Automation in Construction* **2018**, *95*, 155–171.
5. Kumar, S.S.; Abraham, D.M.; Jahanshahi, M.R.; Iseley, T.; Starr, J. Automated Defect Classification in Sewer Closed Circuit Television Inspections Using Deep Convolutional Neural Networks. *Automation in Construction* **2018**, *91*, 273–283.
6. Ding, L.; Goshtasby, A. On the Canny Edge Detector. *Pattern recognition* **2001**, *34*, 721–725.
7. Chen, T.; Wu, Q.H.; Rahmani-Torkaman, R.; Hughes, J. A Pseudo Top-Hat Mathematical Morphological Approach to Edge Detection in Dark Regions. *Pattern Recognition* **2002**, *35*, 199–210.
8. Su, T.-C.; Yang, M.-D.; Wu, T.-C.; Lin, J.-Y. Morphological Segmentation Based on Edge Detection for Sewer Pipe Defects on CCTV Images. *Expert Systems with Applications* **2011**, *38*, 13094–13114.
9. Dong, P. Implementation of Mathematical Morphological Operations for Spatial Data Processing. *Computers & Geosciences* **1997**, *23*, 103–107.
10. Oliveira, H.; Correia, P.L. Automatic Road Crack Segmentation Using Entropy and Image Dynamic Thresholding. In Proceedings of the 2009 17th European signal processing conference; IEEE, 2009; pp. 622–626.
11. Rajeswaran, A.; Narasimhan, S.; Narasimhan, S. A Graph Partitioning Algorithm for Leak Detection in Water Distribution Networks. *Computers & Chemical Engineering* **2018**, *108*, 11–23.
12. Wolf, S.; Pape, C.; Bailoni, A.; Rahaman, N.; Kreshuk, A.; Kothe, U.; Hamprecht, F. The Mutex Watershed: Efficient, Parameter-Free Image Partitioning. In Proceedings of the Proceedings of the European Conference on Computer Vision (ECCV); 2018; pp. 546–562.
13. Almanza-Ortega, N.N.; Flores-Vázquez, J.M.; Martínez-Añorve, H.; Pérez-Ortega, J.; Zavala-Díaz, J.C.; Mexicano-Santoyo, A.; Carmona-Fraustro, J.C. Corrosion Analysis through an Adaptive Preprocessing Strategy Using the K-Means Algorithm. *Procedia Computer Science* **2023**, *219*, 586–595.
14. Kim, B.; Kwon, J.; Choi, S.; Noh, J.; Lee, K.; Yang, J. Corrosion Image Monitoring of Steel Plate by Using K-Means Clustering. *Journal of the Korean institute of surface engineering* **2021**, *54*, 278–284.
15. Jing, J.; Liu, S.; Wang, G.; Zhang, W.; Sun, C. Recent Advances on Image Edge Detection: A Comprehensive Review. *Neurocomputing* **2022**, *503*, 259–271.
16. Wang, M.; Luo, H.; Cheng, J.C. Towards an Automated Condition Assessment Framework of Underground Sewer Pipes Based on Closed-Circuit Television (CCTV) Images. *Tunnelling and Underground Space Technology* **2021**, *110*, 103840.
17. Katsamenis, I.; Protopapadakis, E.; Doulamis, A.; Doulamis, N.; Voulodimos, A. Pixel-Level Corrosion Detection on Metal Constructions by Fusion of Deep Learning Semantic and Contour Segmentation. In *Advances in Visual Computing*; Bebis, G., Yin, Z., Kim, E., Bender, J., Subr, K., Kwon, B.C., Zhao, J., Kalkofen, D., Baciu, G., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, 2020; Vol. 12509, pp. 160–169 ISBN 978-3-030-64555-7.
18. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*; Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, 2015; Vol. 9351, pp. 234–241 ISBN 978-3-319-24573-7.
19. Su, H.; Wang, X.; Han, T.; Wang, Z.; Zhao, Z.; Zhang, P. Research on a U-Net Bridge Crack Identification and Feature-Calculation Methods Based on a CBAM Attention Mechanism. *Buildings* **2022**, *12*, 1561.
20. Ran, Q.; Wang, N.; Wang, X.; He, Z.; Zhao, Y. Improved U-Net Drainage Pipe Defects Detection Algorithm Based on CBMA Attention Mechanism. In Proceedings of the 2024 9th International Conference on Electronic Technology and Information Science (ICETIS); IEEE, 2024; pp. 342–347.
21. Li, Y.; Wang, H.; Dang, L.M.; Piran, M.J.; Moon, H. A Robust Instance Segmentation Framework for Underground Sewer Defect Detection. *Measurement* **2022**, *190*, 110727.

 doi:10.20944/preprints202507.1590.v1

22. Liu, Y.; Bai, X.; Wang, J.; Li, G.; Li, J.; Lv, Z. Image Semantic Segmentation Approach Based on DeepLabV3 plus Network with an Attention Mechanism. *Engineering Applications of Artificial Intelligence* **2024**, *127*, 107260.
23. Wang, N.; Zhang, J.; Song, X. A Pipeline Defect Instance Segmentation System Based on SparseInst. *Sensors* **2023**, *23*, 9019.
24. Forkan, A.R.M.; Kang, Y.-B.; Jayaraman, P.P.; Liao, K.; Kaul, R.; Morgan, G.; Ranjan, R.; Sinha, S. CorrDetector: A Framework for Structural Corrosion Detection from Drone Images Using Ensemble Deep Learning. *Expert Systems with Applications* **2022**, *193*, 116461.
25. Li, Y.; Yang, Y.; Liu, Y.; Zhong, F.; Zheng, H.; Wang, S.; Wang, Z.; Huang, Z. A Novel Method for Semantic Segmentation of Sewer Defects Based on StyleGAN3 and Improved Deeplabv3+. *J Civil Struct Health Monit* **2025**, doi:10.1007/s13349-025-00919-9.
26. Papamarkou, T.; Guy, H.; Kroencke, B.; Miller, J.; Robinette, P.; Schultz, D.; Hinkle, J.; Pullum, L.; Schuman, C.; Renshaw, J. Automated Detection of Corrosion in Used Nuclear Fuel Dry Storage Canisters Using Residual Neural Networks. *Nuclear Engineering and Technology* **2021**, *53*, 657–665.
27. Jin, Q.; Han, Q.; Qian, J.; Sun, L.; Ge, K.; Xia, J. Drainage Pipeline Multi-Defect Segmentation Assisted by Multiple Attention for Sonar Images. *Applied Sciences (2076-3417)* **2025**, *15*.
28. Wang, M.; Cheng, J.C.P. A Unified Convolutional Neural Network Integrated with Conditional Random Field for Pipe Defect Segmentation. *Computer aided Civil Eng* **2020**, *35*, 162–177, doi:10.1111/mice.12481.
29. Malashin, I.; Tynchenko, V.; Nelyub, V.; Borodulin, A.; Gantimurov, A.; Krysko, N.V.; Shchipakov, N.A.; Kozlov, D.M.; Kusyy, A.G.; Martysyuk, D. Deep Learning Approach for Pitting Corrosion Detection in Gas Pipelines. *Sensors* **2024**, *24*, 3563.
30. Das, A.; Dorafshan, S.; Kaabouch, N. Autonomous Image-Based Corrosion Detection in Steel Structures Using Deep Learning. *Sensors* **2024**, *24*, 3630.
31. Du, C.; Wang, K. Drainage Pipeline Defect Detection System Based on Semantic Segmentation. *Symmetry* **2024**, *16*, 1477.
32. Nash, W.T.; Powell, C.J.; Drummond, T.; Birbilis, N. Automated Corrosion Detection Using Crowdsourced Training for Deep Learning. *Corrosion* **2020**, *76*, 135–141.
33. Burton, B.; Nash, W.T.; Birbilis, N. RustSEG -- Automated Segmentation of Corrosion Using Deep Learning 2022.
34. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep High-Resolution Representation Learning for Human Pose Estimation. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2019; pp. 5693–5703.
35. Sun, K.; Zhao, Y.; Jiang, B.; Cheng, T.; Xiao, B.; Liu, D.; Mu, Y.; Wang, X.; Liu, W.; Wang, J. High-Resolution Representations for Labeling Pixels and Regions 2019.
36. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. Cbam: Convolutional Block Attention Module. In Proceedings of the Proceedings of the European conference on computer vision (ECCV); 2018; pp. 3–19.
37. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid Scene Parsing Network. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition; 2017; pp. 2881–2890.
38. Milletari, F.; Navab, N.; Ahmadi, S.-A. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In Proceedings of the 2016 fourth international conference on 3D vision (3DV); Ieee, 2016; pp. 565–571.
39. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In Proceedings of the Proceedings of the IEEE international conference on computer vision; 2017; pp. 2980–2988.