DiffRSD: Diffusion-Based and Integrity-Aware RGB-D Rail Surface Defect Inspection

Zhengyi Liu, Junnan Zhou, Rui Huang, Xianyong Fang, Zhengzheng Tu, and Linbo Wang*

***Abstract*—Rail quality evaluation ensures the safety of railway transportation, where rail surface defect inspection is one of important tasks. Traditional methods adopt encoder-decoder framework, which is difficult to extract discriminative defect features to achieve the integrity of defect inspection. In contrast, we propose DiffRSD, a diffusion-based method which restores the defect mask from a noise conditional on the RGB-D defect image, showing integrity-aware defect inspection ability by the following strategies: (a) superpixel-aware corruption; (b) coarse-to-fine dilation supervision. The first strategy can struggle the model to restore a defect mask from a corrupted mask and make the model rectify the error prediction in the inference stage. The second strategy can locate the defect region in the initial decoding stage and further depict the clear boundary in the later decoding ones. Both strategies improve the performance of defect inspection by experimental verification on NEU RSDDS-AUG RGB-D defect dataset, thus advancing the proposed DiffRSD beyond state-of-the-art methods. The generalization of DiffRSD is further verified by the experiments in RGB rail surface defect inspection dataset and multi-modal pavement crack segmentation dataset. The proposed DiffRSD consistently shows the integrity-aware ability.**[1]

***Index Terms*—rail surface defect inspection, multi-modal, diffusion model**

I. Introduction

During the manufacturing process, rail surface defects such as holes, cracks, and scars can occur due to factors like improper machining processes, excessive thermal stress, insufficient cooling, and other issues [1]. As a result, quality evaluation through defect inspection [2], [3] becomes a critical task in rail intelligent production, ensuring the safety of railway transportation.

Since rail defects are visually similar to their surroundings, depth information has been incorporated into defect inspection [1], [4]–[12]. The results of these studies have confirmed that depth information significantly improves the accuracy of rail surface defect inspection, stimulating the research of RGB-D rail surface defect inspection which combines RGB images with depth images.

This work is supported by National Natural Science Foundation of China under Grant 62376005 (Corresponding author: Linbo Wang).

Zhengyi Liu, Junnan Zhou, Rui Huang, Xianyong Fang, Zhengzheng Tu, and Linbo Wang are with School of Computer Science and Technology, Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, Anhui University, Hefei, China(e-mail: liuzywen@ahu.edu.cn, z253641@163.com, 1615603917@qq.com, fangxianyong@ahu.edu.cn, zhengzhengahu@163.com, wanglb@ahu.edu.cn).

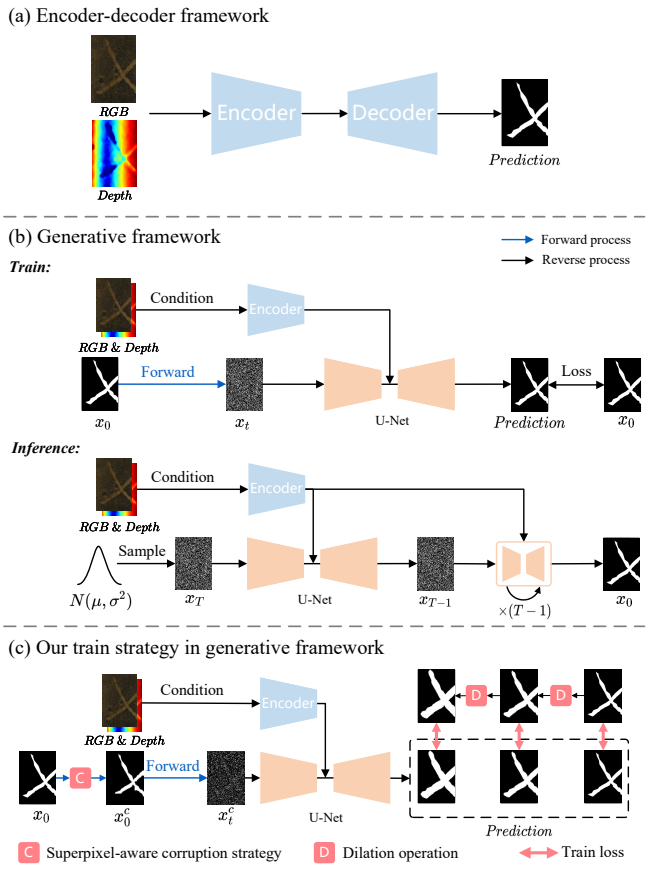


Fig. 1. (a-b) Comparison between encoder-decoder framework and generative framework. (c) Our proposed train strategy in generative framework.

The integrity of inspection is the most important challenge. Existing methods typically follow encoder-decoder framework which extracts features first and then decodes these features into defect masks, as shown in Fig 1 (a). Due to the difficulty in extracting discriminative features that distinguish defects from their surroundings, the integrity of defect inspection remains poor, even with the application of segment anything model (SAM) [13]. Defect masks inspected by SAM are redundant and missing in the first two and last two rows, respectively, as shown in Fig 2.

Instead, a generative framework utilizing denoising diffusion probabilistic models (DDPM) [14] is proposed, as shown in Fig 1 (b). A U-Net is trained to gradually remove the noise from the noised defect mask to a clean mask conditional on RGB-D defect images. Due to the limited capacity of the

trained U-Net to modify the appearance of the defect mask, any incomplete predictions made during the inference step are difficult to correct in later stages. To address the challenge, we propose a superpixel-aware corruption strategy before forward diffusion process, as shown in the "C" module of Fig 1 (c). Defect mask is first corrupted in the appearance before noising, and then attempts to be restored to a normal defect mask with the condition about the RGB-D defect images. The process requires not only denoising but also restoring the mask, which struggling the model generate more discriminative feature which benefits the separation of the defect and its background.

In addition, the different decoding stages of U-Net emphasize on different tasks. In the initial decoding stage, the position of defect is more important to ensure a complete defect region, while in the later stage of decoding, the boundary detail of the defect is progressively depicted. Therefore, we propose a coarse-to-fine dilation supervision strategy to achieve the precise location of defects in the initial decoding stage and boundary clarity in the later decoding stage, as shown in the right part of Fig 1 (c). Defect mask is dilated for the supervision on the U-Net decoder. Specially, dilated defect mask is used to supervise initial decoding process, while original defect mask with clear boundary is responsible for constraining the final result of decoding stages.

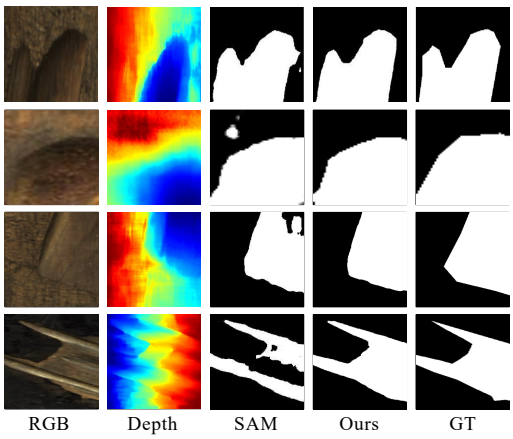


Fig. 2. Integrity comparison between encoder-decoder framework represented by segment anything model (SAM) and our generative method. The first two rows show the redundant inspection, and the last two rows show the missing ones.

The main contributions are summarized as follows:

- A diffusion based generative model, called DiffRSD, is proposed to achieve RGB-D rail surface defect inspection, ensuring the safety of railway transportation. Compared with encoder-decoder framework, the generative model can extract more discriminative features which benefit the separation of defects from surroundings.
- To achieve the integrity of defect inspection, we propose a superpixel-aware corruption strategy before forward diffusion process. The random corruption strategy during the training stage simulates the uncertainty of inference process, enhancing the restore ability from incomplete predictions.
- To ensure the precise location of the defects and the clarity of the boundaries, we propose a coarse-fine dilation supervision strategy. The strategy first supervises the initial decoding result using a dilated ground truth and then restrain the later decoding result using a ground truth with sharp boundary, further achieving the integrity of inspection.

II. Related works

A. RGB-D rail surface defect inspection

Rail surface defect inspection [15] ensures the rail transportation safety by intelligent quality assessment. Depth sensors have been widely applied in computer vision tasks, such as vision navigation [16], scene segmentation [17], road anomaly segmentation [18], semantic segmentation [19], and mirror segmentation [20]. Recently, CLANet [1] introduces the depth modality to defect inspection because depth difference between the defect and normal surface is more prominent compared with color texture information. In the work, NEU RSDDS-AUG industrial RGB-D dataset is constructed and a baseline model with multi-modal fusion and dual stream decoder is also provided. DRERNet [4] achieves the multi-modal fusion both in the encoder and the decoder. FHENet [5] is a light-weight method which incorporates the edge clues for real-time purpose. CSANet [6] exploits bidirectional alignment mechanism between contour and semantic perspectives. MENet [7] integrates the feature quality evaluation into the model and further proposes a knowledge distilled model which can mitigate the parameters of the model. SA2F [8] uses the skeleton images to guide the detection and focuses on the speed of inference. PENet [9] uses knowledge distill to achieve a balance between speed and performance. SAINet [10] combines self-attention and cross-attention to achieve the multi-modal fusion inference. NCLNet [11] leverages the trade-off between the performance and cost via knowledge distillation. DSSNet [12] builds the first semi-supervised method for performance improvement with 20% labelled samples. The aforementioned methods follow encoder-decoder framework. Instead, we propose a diffusion-based method to address the integrity problem of rail surface defect inspection.

B. Diffusion based segmentation

Diffusion based methods have been widely adopted in medical image segmentation. Segmentation mask is gradually noised in forward process and then a neural network is trained to restore the original mask. DMIISE [21], MedSegDiff [22], MedSegDiff-V2 [23], Diff-UNet [24] and MS-SegDiff [25] follow the paradigm to segment medical image. PD-DDPM [26] focuses on accelerating the inference process. DTAN [27] and TGEDiff [28] further introduce text guidance. DBEF-Net [29] and CriDiff [30] address the challenge of the inconsistency between conditional features and diffusion model features. DiffSeg [31] aligns the training with testing of diffusion model via a recycling train. HiDiffSeg [32] adopts coarse segmentation mask and skeleton mask as conditions first and then employs original image as condition, achieving a coarse-to-fine hierarchical diffusion. CCDM [33] extends diffusion-based segmentation to stochastic segmentation. Due to effectiveness, diffusion-based methods have been also applied in

semantic segmentation and depth estimation [34], camouflaged object detection [35], bi-temporal image change detection [36], [37], segmentation refinement [38], dense matching [39]. To reduce computation cost, DiffusionEdge [40] and Marigold [41] exploit the latent space diffusion process in edge detection and depth estimation, respectively.

To achieve the integrity inspection of rail surface defects, we propose superpixel-aware corruption before forward diffusion process and coarse-to-fine dilation supervision in reverse train stage based on diffusion generative framework.

III. Method

A. Revisiting denoising diffusion probabilistic models

Denoising diffusion probabilistic models (DDPM) comprises two processes: a forward diffusion process and a reverse denoising process. The forward process adds random noise to data, while the reverse process constructs desired data samples from the noise.

Given a sample x_0, the posterior $q(x_{1:T}|x_0)$ is fixed to a Markov chain.

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}) \tag{1}$$

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}), \quad t \in \{1, \cdots T\} \tag{2}$$

where the present state x_t in t time step is noised from the previous state x_{t-1}, β_t is noise schedule, $\mathbf{I}$ is the identity matrix, and $\mathcal{N}$ represents the Gaussian distribution. Any step x_t may be sampled directly from x_0 without the need to generate the intermediate steps.

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I}) \tag{3}$$

where $\bar{\alpha}_t = \prod_{i=1}^{t}(1-\beta_i)$. Using reparameterization trick [42], x_t can be represented as:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}) \tag{4}$$

The forward process can be shown in the blue flow of Fig 1 (b).

The reverse process aims at restoring the original data x_0 by reversing the noising process.

$$p_\theta(x_{0:T-1}|x_T) = \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t) \tag{5}$$

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2\mathbf{I}), \quad t \in \{1, \cdots T\} \tag{6}$$

where mean and variance are respectively defined as:

$$\mu_\theta(x_t, t) = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}x_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t \tag{7}$$

$$\sigma_t^2 = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t \tag{8}$$

DDPM has two kinds of predictions: ϵ-prediction and x_0-prediction.

In the ϵ-prediction, DDPM uses a U-Net neural network $\mathcal{F}_\theta$ to predict the t-step noise ϵ_t, i.e., $\epsilon_t = \mathcal{F}_\theta(x_t, t, c)$, where c is condition. According to the relationship between x_0 and x_t shown in formula 4, the mean in formula 7 can be deduced as:

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_t) \tag{9}$$

To train $\mathcal{F}_\theta$, DDPM randomly picks a clean image x_0 from the dataset and samples a noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$, then picks a random time-step t and updates the network parameters θ in $\mathcal{F}_\theta$ with the following gradient descent step.

$$\bigtriangledown_\theta \left\|\epsilon - \mathcal{F}_\theta(x_t, t, c)\right\|_2^2 \tag{10}$$

During inference, reverse denoising steps are performed starting from a randomly sampled Gaussian noise $x_T \sim \mathcal{N}(0, \mathbf{I})$. For each step t, x_{t-1} is inferred by reparameterization.

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\mathcal{F}_\theta(x_t, t, c)\right) + \sigma_t z \tag{11}$$

where $z \sim \mathcal{N}(0, \mathbf{I})$ when $t > 1$ and $z = 0$ when $t = 1$.

In the x_0-prediction, DDPM trains a U-Net neural network $\mathcal{F}_\theta$ to predict the original input x_0, as shown in 1 (b). Then, the learning objective of U-Net is revised as:

$$\bigtriangledown_\theta \left\|x_0 - \mathcal{F}_\theta(x_t, t, c)\right\|_2^2 \tag{12}$$

According to formula 7–8 and reparameterization, in the inference stage x_{t-1} can be iteratively inferred as:

$$x_{t-1} = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}x_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t + \sigma_t z \tag{13}$$

B. Our model

1) Model training and inference: Based on x_0-prediction manner of DDPM, our model includes training and inference stages.

In the training stage, as shown in Fig 3, a superpixel-aware corruption strategy is proposed to generate the corrupted mask, then a diffusion process is conducted to generate the noised mask. Subsequently, the noised mask is fed into the denoising U-Net conditional on RGB-D image to yield the multi-scale defect prediction, which are supervised by the coarse-to-fine dilation ground truth. The superpixel-aware corruption strategy struggles the model to restore the original mask from a corrupted mask, which indirectly improving the discrimination of defect features from their context. The coarse-to-fine dilation supervision strategy ensures the accuracy of defect masks by emphasizing accurate position in the early decoding stage and sharp boundary in the later ones.

The training algorithm can be seen in Alg. 1. In row 2-5, time step t, train sample pair (I, D) and the corresponding defect mask x_0, and standard Gaussian noise ϵ are respectively sampled. In row 6, the defect mask x_0 is first converted into the corrupted mask x_0^c based on the RGB image I, where $C(\cdot)$ refers to the superpixel-aware corruption. Then in row 7, the corrupted mask is incrementally added with Gaussian noise through a series of steps t to yield x_t^c. In row 8, a denoising U-Net $\mathcal{F}_\theta$ is used to predict the denoised multi-scale mask

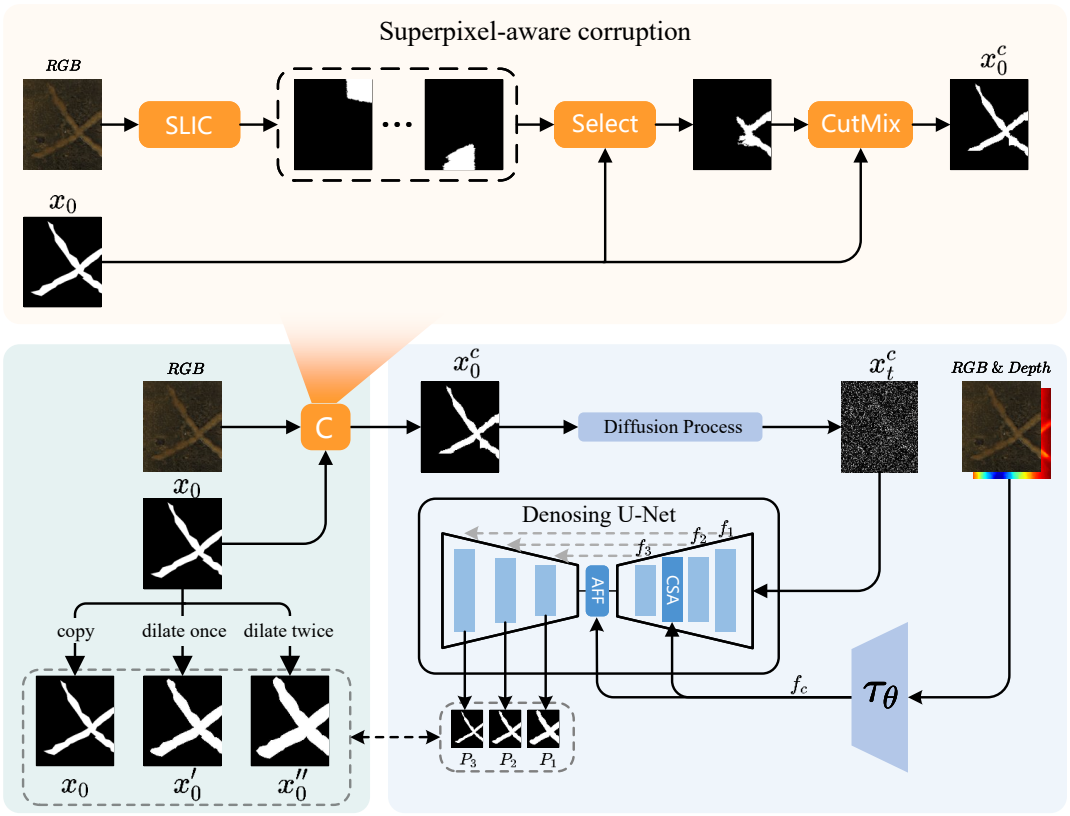


Fig. 3. The framework of DiffRSD. The top part in light orange box is superpixel-aware corruption; the bottom part in the light green box is coarse-to-fine dilation supervision; the bottom part in the light purple box is denoising U-Net.

P_1, P_2, P_3 conditional on time step t and RGB-D image pair (I, D). In rows 9-12, the denoised multi-scale mask P_1, P_2, P_3 are supervised by second-order dilation mask x_0'', first-order dilation mask x_0', and original mask x_0, respectively, where the loss $\mathcal{L}$ includes weighted intersection-over-union (IoU) loss and weighted binary cross entropy (BCE) loss [35]. The model is trained until convergence.

Algorithm 1 Training.

Input: RGB-D image pair (I, D) and defect mask x_0.

1: **repeat**
2: $t \sim Uniform(\{1, \cdots, T\})$
3: $(I, D) \sim q(I, D)$
4: $x_0 \sim q(x_0|I, D)$
5: $\epsilon \sim \mathcal{N}(0, \mathbf{I})$
6: $x_0^c = C(x_0, I)$
7: $x_t^c = \sqrt{\bar{\alpha}_t} x_0^c + \sqrt{1-\bar{\alpha}_t}\epsilon$
8: $\{P_1, P_2, P_3\} = \mathcal{F}_\theta(x_t^c, t, I, D)$
9: $x_0' = dilate(x_0)$
10: $x_0'' = dilate(dilate(x_0))$
11: Take gradient descent step on
12: $\nabla_\theta \mathcal{L}(P_1, x_0'') + \mathcal{L}(P_2, x_0') + \mathcal{L}(P_3, x_0)$;
13: **until** converge.

In the inference stage, to improve the efficiency, we use fewer time steps with the interval δ to skip-step infer the final defect mask image x_0 from Gaussian noise $x_T \sim \mathcal{N}(0, \mathbf{I})$

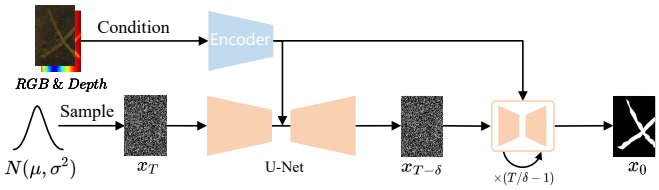


Fig. 4. The skip-step inference.

following DDIM [43], shown in Fig 4. Specifically, as shown in Alg. 2, the denoising U-Net is responsible to predict P_3 from x_t ($t \in \{T, T-\delta, \cdots\}$), then according to formula 13, x_{t-1} can be represented as:

$$x_{t-\delta} = \frac{\sqrt{\bar{\alpha}_{t-\delta}}\beta_t}{1-\bar{\alpha}_t} P_3 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-\delta})}{1-\bar{\alpha}_t} x_t + \sigma_t z \qquad (14)$$

where $x_{t-\delta}$ is the inference in t-δ time step, and $\bar{\alpha}_{t-\delta} = \prod_{i=1}^{t-\delta}(1-\beta_i)$. The final result is obtained through skip-step inference.

Next, we will elaborate on superpixel-aware corruption strategy, denoising U-Net, and coarse-to-fine dilation supervision.

2) Superpixel-aware corruption strategy: The denoising U-Net is used to predict P_3, and iteratively infer $x_{t-\delta}$ from x_t until x_0 using formula 14. However, P_3 is a predicted result which may be inaccurate so that the inferred $x_{t-\delta}$ is also inaccurate. Once an error is made in a step of inference, this error is difficult to be rectified at a later stage. Therefore, to

Algorithm 2 Inference.

Input: RGB-D image pair (I, D).

$x_T \sim \mathcal{N}(0, \mathbf{I})$
for $t = T, T-\delta, T-2\delta, \cdots$ **do**
  $P_3 = \mathcal{F}_\theta(x_t, t, I, D)$
  if $t > 1$ **then**
    $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$
  else
    $\mathbf{z}$=0
  end if
  $x_{t-\delta} = \frac{\sqrt{\bar{\alpha}_{t-\delta}}\beta_t}{1-\bar{\alpha}_t} P_3 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-\delta})}{1-\bar{\alpha}_t} x_t + \sigma_t \mathbf{z}$
end for
return x_0.

equip the model with the ability to correct prediction errors, we propose a superpixel-aware corruption strategy before forward diffusion process to simulate the inference scene in the train phase and align data distribution of train and inference.

Superpixel-aware corruption includes three steps. At first, the RGB image is segmented into multiple superpixels via SLIC method [44]. Then, a superpixel that has the largest intersection with original mask is selected. Last, CutMix operation between original mask and selected superpixel is conducted to generate the corrupted mask.

Specially, as shown in the top of Fig 3, before the forward diffusion process, we first segment RGB image I into multiple superpixels via SLIC.

$$\{S_i\}_{i=1}^{N} = \mathit{SLIC}(I) \tag{15}$$

where the number of superpixels N is 20 with the compactness of 10. To minimize the subsequent corruption and reduce computational complexity, depth image is not involved.

In all the superpixels $\{S_i\}_{i=1}^{N}$, a superpixel that has the largest IoU value with defect mask x_0 is selected.

$$S = \mathit{Select}(\{S_i\}_{i=1}^{N}, x_0) \tag{16}$$

where $\mathit{Select}(\cdot)$ will pick out a superpixel S which is most similar with defect mask x_0.

Next, also aimed at achieving minimal corruption, CutMix operation between defect mask x_0 and selected superpixel S is conducted to generate the corrupted defect mask x_0^c.

$$x_0^c = \mathit{CutMix}(S, x_0) = \mathit{Mask} \odot S + (1 - \mathit{Mask}) \odot x_0 \tag{17}$$

where *Mask* denotes a binary mask with the size of half of image area, random position, and random width and height, indicating where to drop out and fill in from two images [45].

Last, the corrupted defect mask x_0^c will be noised to yield t-step noised corrupted defect mask x_t^c.

$$x_t^c = \sqrt{\bar{\alpha}_t} x_0^c + \sqrt{1 - \bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}) \tag{18}$$

3) Denoising U-Net: Denoising U-Net, abbreviated as U-Net, takes the t-step noised corrupted defect mask x_t^c and RGB-D image pair (I, D) as the input, and outputs the denoised multi-scale mask P_1, P_2, P_3, which is illustrated in lower right of Fig 3.

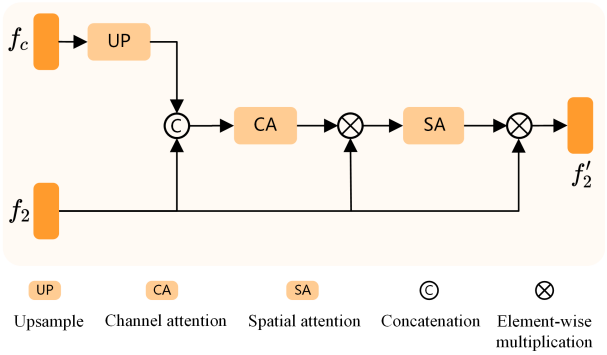


Fig. 5. Channel and spatial attention (CSA).

To instruct U-Net to generate a defect mask corresponding to defect image, RGB-D defect image pair (I, D) is first fed into an encoder to generate the conditional feature f_c.

$$f_c = \mathcal{T}_\theta(I, D) \tag{19}$$

where $\mathcal{T}_\theta$ adopts pretrained DFormer encoder [46] which outputs concatenated RGB features from the last three stages.

Next, the conditional feature f_c is injected into U-Net encoder after the second layer by a channel and spatial attention (CSA) [47].

$$f_2' = \mathit{CSA}(f_2, f_c) \tag{20}$$

where f_2' updates the second layer U-Net feature f_2.

The details of CSA is shown in Fig 5. Specifically, conditional feature f_c is concatenated with U-Net feature f_2, and fed into a channel attention module to get attentive channel mask, which is used to enhance U-Net feature in a channel manner. Next, enhanced U-Net feature is fed into a spatial attention module again to generate attentive spatial mask, which is used to enhance the U-Net feature in a spatial manner.

$$f_2' = f_2 \times \mathit{SA}(f_2 \times \mathit{CA}(\mathit{Cat}(\mathit{UP}(f_c), f_2))) \tag{21}$$

where $\mathit{UP}(\cdot)$ is upsampling operation, $\mathit{Cat}(\cdot)$ denotes concatenation, $\mathit{CA}(\cdot)$ and $\mathit{SA}(\cdot)$ are channel and spatial attention operation which are proposed by CBAM [48], "$\times$" is element-wise multiplication operation.

It is well known that the convolution operation in U-Net is more concerned with local high frequency details and the transformer block in DFormer is more concerned with global low frequency information. To dynamically adjust the ratio of low-frequency and high-frequency information and properly fuse the low-frequency and high-frequency information, the conditional feature f_c obtained by the DFormer encoder need to be fused with the U-Net feature obtained in the highest layer by an adaptive FFT-filter (AFF) [40].

$$f_3' = \mathit{AFF}(f_3, f_c) \tag{22}$$

where f_3' updates f_3 in U-Net and will be fed into U-Net decoder.

The details of AFF is shown in shown in Fig 6. Specifically, U-Net feature f_3 and conditional feature f_c are converted into Fourier space by 2D Fast Fourier Transform (FFT), respectively. A pair of learnable weight maps W_3 and W_c

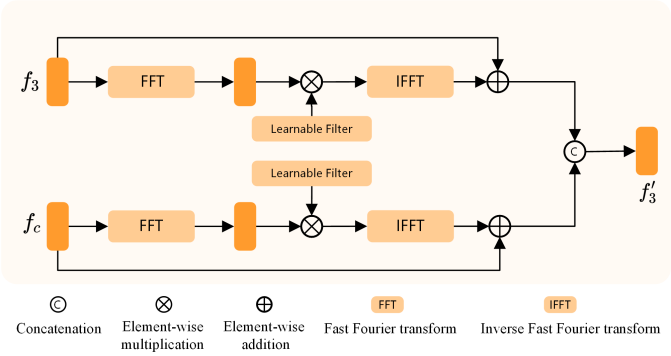


Fig. 6. Adaptive FFT-filter (AFF).

which are regarded as adaptive spectrum filters are learnt and multiply f_3 and f_c, respectively. After filtering out the useless components adaptively, the features from the frequency domain are projected back to the spatial domain by Inverse Fast Fourier Transform (IFFT). Next, a residual connection is adopted to avoid filtering out useful information.

$$q_3 = f_3 + IFFT(W_3 \times FFT(f_3)) \tag{23}$$

$$q_c = f_c + IFFT(W_c \times FFT(f_c)) \tag{24}$$

Both features are concatenated to obtain the optimal features.

$$f_3' = Cat(q_3, q_c) \tag{25}$$

Finally, U-Net decoder outputs the denoised multi-scale mask P_1, P_2 and P_3 progressively.

4) Coarse-to-fine dilation supervision: Defects such as cracks, scratches, and dents often resemble the surrounding surface in terms of color, texture, and other visual features, making it challenging to distinguish them from their surroundings. Allowing some tolerance for edge details in the early stages of decoding enables the model to first identify the approximate location and then refine the details, ensuring both defect completeness and boundary clarity. Therefore, we propose a coarse-to-fine dilation supervision method.

Specially, the mask x_0 is first dilated 8 times and 16 times to generate a first-order dilation mask x_0' and a second-order dilation mask x_0'', respectively.

$$\begin{cases} x_0' = dilate(x_0) \\ x_0'' = dilate(dilate(x_0)) \end{cases} \tag{26}$$

where $dilate(\cdot)$ conducts the dilation operation with the number of 8.

The dilation and original masks are used to supervise the output of U-Net decoder $\{P_1, P_2, P_3\}$. Note that the three outputs need to be resized to the same size with the ground truth by upsampling operation. The low-resolution feature P_1 in the initial decoding process is supervised by second-order dilation mask x_0'', which emphasizing on the scope of defects. The middle-resolution feature P_2 in the middle decoding process is supervised by first-order dilation mask x', which progressively focusing on the boundary of defects. The high-resolution feature P_3 in the later decoding process is supervised by the original mask x_0, which further refining the boundary which is critical for defect inspection because the challenge of defect inspection lies on similar foreground and background.

IV. Experiments

A. Datasets and evaluation metrics

1) Dataset: NEU RSDDS-AUG [1] contains 1,500 training RGB-D samples and 362 testing ones.

2) Evaluation metrics: The evaluation metrics are used to estimate the performance, including S-measure (S_m) [49], mean E-measure ($meanE_m$) [50], mean F-measure ($meanF_m$) [51], and mean absolute error (MAE) [52]. Specifically, the S-measure can evaluate both region-aware and object-aware structural similarity between defect mask and ground truth. The E-measure simultaneously captures global statistics and local pixel matching information. The F-measure is the weighted harmonic mean of precision and recall, which can evaluate the overall performance. The MAE measures the average of the per-pixel absolute difference between the defect mask and the ground truth. The lower the value of MAE and the higher the value of others, the better the performance.

B. Implementation details

The code is implemented in Pytorch toolbox on a PC with an NVIDIA RTX 3090 GPU. During training, each image is uniformly resized to 352×352. The batch size is 10 and the initial learning rate is 5e-5. The number of training iterations T is 1000, while the number of inference iterations is optional from 10 and 1 for high performance and high efficiency where time step interval δ=100 and δ=1000, respectively. The model converges around 200 epochs by Adam optimizer [53]. The training time is about 20 hours. The model has 286 million parameters, requires 65.8 FLOPs per inference, and achieves a speed of 14.3 frames per second (FPS) when the number of inference iterations is 1. An SNR-based variance schedule proposed in simple diffusion [54] is adopted in diffusion process.

C. Comparison with state-of-the-art methods

1) Quantitative comparisons: At first, our method is compared with state-of-the-art defect inspection methods, including CLANet [1], DRERNet [4], FHENet [5], CSANet [6], MENet [7], SA2F [8], PENet [9], SAINet [10], NCLNet [11], DSSNet [12]. All the metrics of the competitors use the results reported by the original papers. From the results in Table I, we can observe that our method outperforms the others and achieves a significant improvement in all the evaluation metrics.

Secondly, our method is compared with encoder-decoder frameworks represented by DFormer [46] and SAM [13]. The results are shown in Table II. "DFormer E-D" refers to the DFormer encoder-decoder model. Although the SAM only supports the input with RGB image, two SAM encoders with finetuned adapters [55] are adopted in "SAM+Adapter E-D". The results show the worse performance of encoder-decoder

TABLE I
Quantitative comparisons with RGB-D rail surface defect inspection methods.

Methods	Source	$S_m \uparrow$	$meanE_m \uparrow$	$meanF_m \uparrow$	$MAE \downarrow$
CLANet [1]	TMech22	.834	.912	.863	.069
DRERNet [4]	SPL22	.844	.929	.878	.059
FHENet [5]	TIM23	.836	.921	.869	.064
CSANet [6]	SPL23	.861	.928	.885	.058
MENet [7]	TCSVT23	.856	.931	.884	.057
SA2F [8]	TII23	.845	.923	.870	.061
PENet [9]	TIM23	.856	.938	.902	.058
SAINet [10]	OLEN24	.849	.934	.886	.055
NCLNet [11]	TITS24	.860	.941	.903	.055
DSSNet [12]	TITS24	.866	.942	.897	.049
Ours	-	**.873**	**.947**	**.905**	**.045**

framework compared with our diffusion-based and integrity-aware method, which benefits from the restore ability from error and defect inspection integrity constraints.

TABLE II
Quantitative comparison with encoder-decoder framework represented by DFormer and SAM encoder-decoder models.

Methods	$S_m \uparrow$	$meanE_m \uparrow$	$meanF_m \uparrow$	$MAE \downarrow$
DFormer E-D	.854	.937	.888	.053
SAM+Adpater E-D	.858	.936	.889	.053
Ours	**.873**	**.947**	**.905**	**.045**

Thirdly, our method is compared with diffusion-based segmentation models, including MedSegDiff-v2 [23] and CamoDiffusion [35]. Since the two models are RGB based methods, we concatenate the RGB and depth in the input level to better reproduce the corresponding method. The results in Table III show the better performance of our model, which benefits from the restore ability from error and defect inspection integrity constraints.

TABLE III
Quantitative comparison with diffusion-based segmentation models.

Methods	$S_m \uparrow$	$meanE_m \uparrow$	$meanF_m \uparrow$	$MAE \downarrow$
MedSegDiff-v2 [23]	.851	.924	.876	.057
CamoDiffusion [35]	.855	.934	.888	.055
Ours	**.873**	**.947**	**.905**	**.045**

Fourthly, our method is compared with the method without depth input. From the results shown in Table IV, we can observe that the depth images play an auxiliary role in improve the accuracy of rail surface defect inspection.

TABLE IV
Quantitative comparison with the method without depth input.

Methods	$S_m \uparrow$	$meanE_m \uparrow$	$meanF_m \uparrow$	$MAE \downarrow$
w/o depth	.861	.939	.897	.050
Ours w/ depth	**.873**	**.947**	**.905**	**.045**

2) Qualitative comparisons with defect inspection methods: Fig 7 further shows the qualitative comparison among RGB-D rail surface defect inspection methods and encoder-decoder models. Our method produces results that are closer to ground truth, achieving the completeness of the inspection and the clarity of the boundaries. In 1^{st}-3^{rd} rows, our method successfully detect the defects from similar background, while the detection result of the comparison methods are poor. In 4^{th}-6^{th} rows, our method is not affected by background noise, the results are the same with ground truth, demonstrating the model's ability to recover from errors. In 7^{th}-9^{th} rows, our method generates fine-grained results due to refinement ability of improved diffusion models. However, our method is poor in some challenging scenes as shown in Fig 8. In the 1^{st}-2^{nd} rows, due to lighting effects, it is difficult to distinguish defects from the background. In the 3^{rd}-4^{th} rows, distorted depth information interferes with the quality of defect segmentation. The inherent defects of color images and depth maps increase the difficulty of inspection.

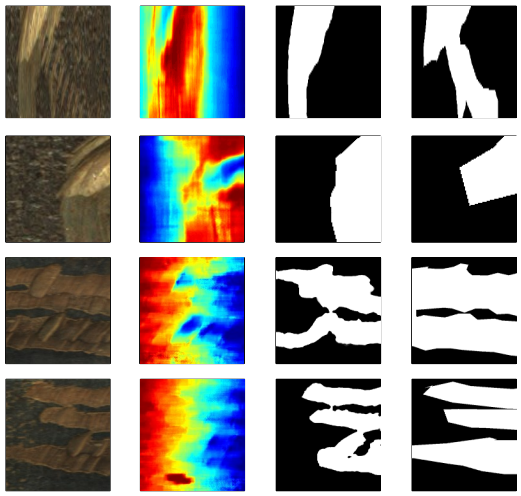


Fig. 8. Failure cases.

D. Ablation studies

TABLE V
Ablation study on CSA and AFF module, superpixel-aware corruption, and coarse-to-fine dilation supervision.

DDPM	CSA+AFF	Corruption	Supervision	$S_m \uparrow$	$meanE_m \uparrow$	$meanF_m \uparrow$	$MAE \downarrow$
✓				.854	.933	.880	.053
✓	✓			.859	.937	.889	.052
✓	✓	✓		.866	.943	.898	.048
✓	✓		✓	.865	.942	.899	.050
✓	✓	✓	✓	**.873**	**.947**	**.905**	**.045**

1) Effectiveness analysis of CSA and AFF in denoising U-Net: By comparing the first and second rows of Table V, we can observe that the performance of denoising U-Net is improved by adding CSA and AFF modules on basic DDPM, verifying the effectiveness of the two modules. CSA focuses on the important information by attention mechanism, and AFF helps in removing noise while preserving useful edge information by the learned weights adaptively tuned to the target distribution of different frequencies.

2) Effectiveness analysis of superpixel-aware corruption: By comparing the second and third rows of Table V, we can observe the effectiveness of superpixel-aware corruption. The performance is significantly improved when adding

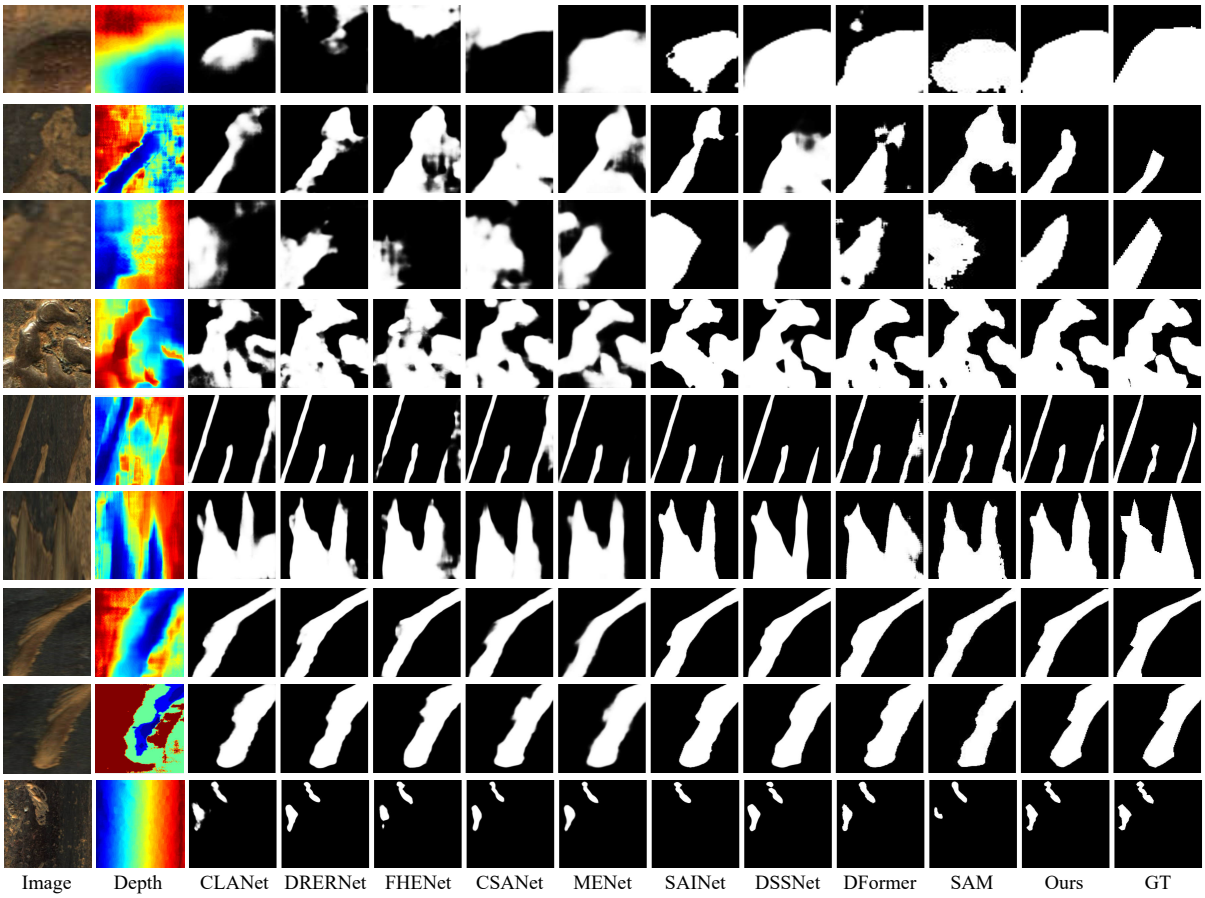


Fig. 7. Visual comparison of RGB-D rail surface defect inspection methods.

superpixel-aware corruption strategy. Four evaluation metrics are improved by 0.007, 0.006, 0.009, and 0.004, respectively. Moreover, by comparing the fourth and fifth rows of Table V, we can also observe the effectiveness of superpixel-aware corruption when using coarse-to-fine dilation supervision. The visualization of inference process in Fig 9 provides evidence that our method can recover from error defect inspection indicated by the red boxes, because our diffusion model is trained to restore the original mask from a corrupted mask. To achieve the goal, the model is struggled to extract the more discriminative features from RGB-D image pair.

3) Effectiveness analysis of coarse-to-fine dilation supervision: By comparing the second and fourth rows of Table V, we can observe the effectiveness of coarse-to-fine dilation supervision. When adding coarse-to-fine dilation supervision, the defect inspection accuracy is raised by 0.006, 0.005, 0.010, and 0.002, respectively. Moreover, by comparing the third and fifth rows of Table V, we can also observe the effectiveness of coarse-to-fine dilation supervision when using superpixel-aware corruption. Fig 10 further gives the effectiveness evidence of coarse-to-fine dilation supervision. The initial output P_1 of U-Net locates the coarse region of defects, while the later output P_3 emphasizes more on the sharp boundary. In contrast, the model with single supervision in the end of the decoder instead of coarse-to-fine dilation supervision generates the defect mask far away from ground truth. Table VI further compares multi-scale supervision and our coarse-to-fine dilation supervision. Our supervision allows for greater tolerance in the early stages of decoding, emphasizing positional accuracy over fine details, and only focuses on refining edge details in the later stages, resulting in better overall performance.

TABLE VI
Comparison with multi-scale supervision.

Methods	$S_m \uparrow$	$meanE_m \uparrow$	$meanF_m \uparrow$	$MAE \downarrow$
Multi-scale supervision	.869	.943	.900	.047
Ours	**.873**	**.947**	**.905**	**.045**

4) Analysis of some parameter settings: In the superpixel-aware corruption strategy, the impact on performance about the size of CutMix mask between defect mask and selected superpixel is analyzed in Table VII. According to the experimental results, the size of the mask is set as 1/2 of the image. The radio is 0 indicating no corruption strategy.

TABLE VII
Analysis about the size of Cutmix mask.

Radio	$S_m \uparrow$	$meanE_m \uparrow$	$meanF_m \uparrow$	$MAE \downarrow$
0	.865	.942	.899	.050
1/3	.866	.943	.903	.048
1/2 (Ours)	**.873**	**.947**	**.905**	**.045**
2/3	.870	.946	.903	.046

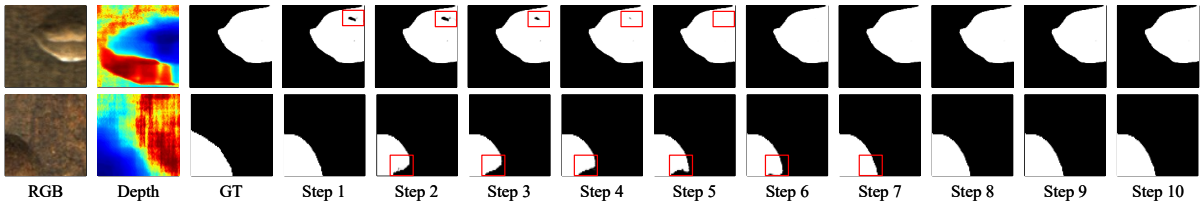


Fig. 9. The inference process indicating the ability to recover from error defect inspection of our method. The red boxes represent the error inspection region.

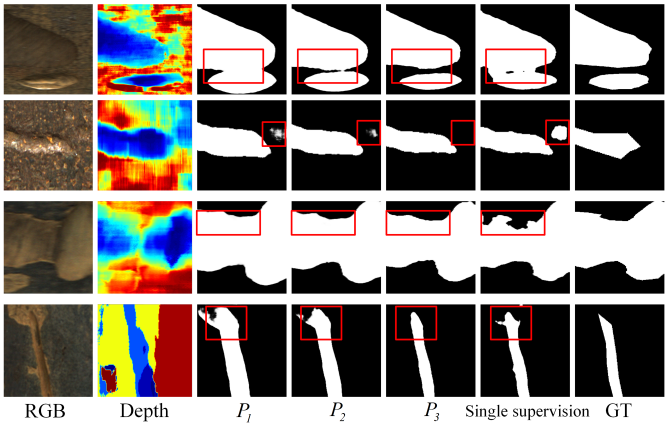


Fig. 10. P_1, P_2, and P_3 show the decoding process indicating the refinement ability from coarse to fine of our method. "Single supervision" denotes the performance of the model with single supervision in the end of decoder. The red boxes represent the refinement details.

In the coarse-to-fine dilation supervision, the impact on performance about the number of dilations in $dilate(\cdot)$ operation is analyzed in Table VIII. According to the experimental results, the number of first-order dilation is 8 and the number of second-order dilation is 16.

TABLE VIII
Analysis on the number of first-order dilation and second-order dilation.

The number of dilation	$S_m \uparrow$	$meanE_m \uparrow$	$meanF_m \uparrow$	$MAE \downarrow$
(16,32)	.870	.945	.902	.046
(8,16)(Ours)	**.873**	**.947**	**.905**	**.045**
(4,8)	.871	.945	.904	.047

In the inference process, the impact on performance about inference time steps is analyzed in Table IX. According to the experimental results, when the inference time step is 1, the performance experiences only a slight decline, while the FPS is improved tenfold. Accordingly, an inference time step of 10 is used when performance is prioritized, while a time step of 1 is chosen when time efficiency is the main concern.

TABLE IX
Analysis about the inference time step.

Steps	FPS	$S_m \uparrow$	$meanE_m \uparrow$	$meanF_m \uparrow$	$MAE \downarrow$
1	14.3	.871	.946	.904	**.045**
10 (Ours)	1.4	**.873**	**.947**	.905	**.045**
20	0.8	**.873**	**.947**	**.906**	**.045**

E. Generalizability Validation

The proposed DiffRSD is designed for RGB-D rail surface defect inspection. To verify the generalization of our method, we apply the DiffRSD network to RGB rail surface defect inspection. RGB images are copied to meet the demand of dual-modal input. Following the settings in SA2F [8], NRSD-MN dataset [56], which consists of 2,086 training samples, 885 validating samples, and 965 testing ones, is adopted to evaluate the performance of compared methods. Table X and Fig 11 give the quantitative and qualitative comparison with MCnet [56], SA2F [8], and ETD [57]. The results verify the superior of the diffusion-based and integrity-aware DiffRSD over encoder-decoder framework in rail surface defect inspection task, achieving more excellent performance in inspection integrity.

TABLE X
Quantitative comparisons with RGB rail surface defect inspection methods.

Methods	Source	$S_m \uparrow$	$meanE_m \uparrow$	$meanF_m \uparrow$	$MAE \downarrow$
MCnet [56]	TIM21	.862	.955	.845	.018
SA2F [8]	TII23	.872	.936	.835	.023
ETD [57]	OLEN24	.877	.949	.824	.022
Ours	-	**.886**	**.960**	**.858**	**.016**

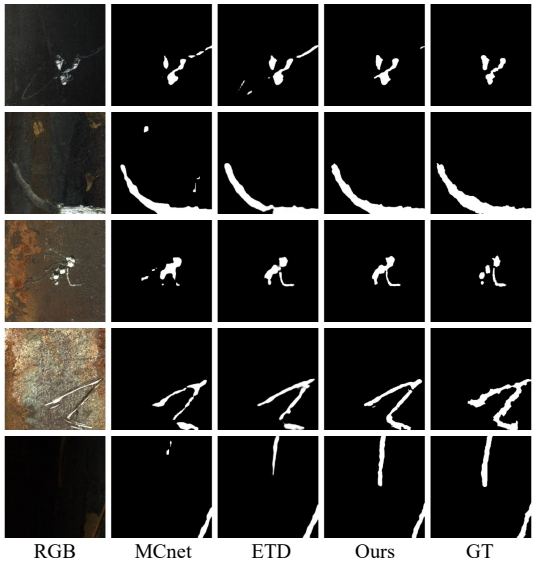


Fig. 11. Qualitative comparison with RGB rail surface defect inspection methods.

Furthermore, the proposed method is applied to multi-modal pavement crack segmentation, using image pairs as

input, which include intensity images from RGB cameras and range images from three-dimensional sensors. The FIND [58] dataset is collected from multiple pavements under varying ambient environments and highly complex surface conditions using a laser scanning device mounted on a survey vehicle. Specifically, it contains 2,500 pairs of intensity and range image patches, each with a resolution of 256×256 pixels. We manually split the data, using 80% for training and 20% for testing. The comparison results in Table XI verify the generalizability of the proposed method in the scenarios where the background and foreground are similar and refinement of the results is required.

TABLE XI
Generalizability experiment on multi-modal pavement crack segmentation.

Methods	IoU	F1	Precision	Recall
Crack Transformer [59]	78.82	88.16	88.89	87.44
SegFormer [60]	77.58	87.37	89.86	85.02
Swin-UperNet [61]	77.71	87.46	87.90	87.02
UNet-FCN [62]	79.27	88.44	89.96	86.97
VGG19-FCN [63]	75.62	86.12	82.34	90.26
CrackFusionNet [64]	81.64	89.89	89.72	90.06
Ours	**83.80**	**92.02**	**93.25**	**90.82**

V. Conclusions

To address the integrity challenge of rail surface defect inspection, we propose a diffusion-based and integrity-aware model. Based on DDPM framework, we add superpixel-aware corruption strategy to infer the defect mask from a corrupted mask, ensuring the completeness of defect inspection. Moreover, we add coarse-to-fine dilation supervision to both locate the scope of defect in the early decoding and depict the sharp boundary in the later decoding.

Diffusion-based methods gradually refine defect masks, but this process incurs significant time costs. As a result, the inference speed is relatively slow and cannot meet real-time requirements. To address this issue, future work could explore lightweight diffusion architectures, accelerated denoising schedulers, or hybrid frameworks that combine the strengths of diffusion and traditional encoder-decoder models to balance accuracy and efficiency.

References

[1] J. Wang, K. Song, D. Zhang, M. Niu, and Y. Yan, "Collaborative learning attention network based on RGB image and depth image for surface defect inspection of no-service rail," *T-Mech*, vol. 27, no. 6, pp. 4874–4884, 2022.

[2] H. Zhang, Y. Song, Y. Chen, H. Zhong, L. Liu, Y. Wang, T. Akilan, and Q. J. Wu, "MRSDI-CNN: Multi-model rail surface defect inspection system based on convolutional neural networks," *TITS*, vol. 23, no. 8, pp. 11 162–11 177, 2021.

[3] X. Ni, H. Liu, Z. Ma, C. Wang, and J. Liu, "Detection for rail surface defects via partitioned edge feature," *TITS*, vol. 23, no. 6, pp. 5806–5822, 2021.

[4] J. Wu, W. Zhou, W. Qiu, and L. Yu, "Depth repeated-enhancement RGB network for rail surface defect inspection," *SPL*, vol. 29, pp. 2053–2057, 2022.

[5] W. Zhou and J. Hong, "FHENet: Lightweight feature hierarchical exploration network for real-time rail surface defect inspection in RGB-D images," *TIM*, vol. 72, pp. 1–8, 2023.

[6] J. Yang, W. Zhou, R. Wu, and M. Fang, "CSANet: Contour and semantic feature alignment fusion network for rail surface defect detection," *SPL*, pp. 972–976, 2023.

[7] W. Zhou, J. Hong, W. Yan, and Q. Jiang, "Modal evaluation network via knowledge distillation for no-service rail surface defect detection," *TCSVT*, pp. 3930–3942, 2023.

[8] L. Huang and A. Gong, "Surface defect detection for no-service rails with skeleton-aware accurate and fast network," *TII*, pp. 4571–4581, 2023.

[9] B. Wang, W. Zhou, W. Yan, Q. Jiang, and R. Cong, "PENet-KD: Progressive enhancement network via knowledge distillation for rail surface defect detection," *TIM*, pp. 1–11, 2023.

[10] Y. Yan, X. Jia, K. Song, W. Cui, Y. Zhao, C. Liu, and J. Guo, "Specificity autocorrelation integration network for surface defect detection of no-service rail," *OPT LASER ENG*, vol. 172, pp. 1–10, 2024.

[11] X. Sun, W. Zhou, and X. Qian, "Normalized cyclic loop network for rail surface defect detection using knowledge distillation," *TITS*, pp. 16 561–16 573, 2024.

[12] J. Wang, G. Li, G. Qiu, G. Ma, J. Xi, and N. Yu, "Depth-assisted semi-supervised RGB-D rail surface defect inspection," *TITS*, pp. 8042–8052, 2024.

[13] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollar, and R. Girshick, "Segment anything," in *ICCV*, 2023, pp. 4015–4026.

[14] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *NeurIPs*, vol. 33, pp. 6840–6851, 2020.

[15] F. Guo, J. Liu, Y. Qian, and Q. Xie, "Rail surface defect detection using a transformer-based network," *JIII*, vol. 38, pp. 1–13, 2024.

[16] C. Sun, X. Wu, J. Sun, C. Sun, M. Xu, and Q. Ge, "Saliency-induced moving object detection for robust RGB-D vision navigation under complex dynamic environments," *TITS*, vol. 24, no. 10, pp. 10 716–10 734, 2023.

[17] L. Li, B. Qian, J. Lian, W. Zheng, and Y. Zhou, "Traffic scene segmentation based on RGB-D image and deep learning," *TITS*, vol. 19, no. 5, pp. 1664–1669, 2017.

[18] B. Ganguly, D. Dey, and S. Munshi, "An unsupervised learning approach for road anomaly segmentation using RGB-D sensor for advanced driver assistance system," *TITS*, vol. 23, no. 10, pp. 19 042–19 053, 2022.

[19] Y. Qian, L. Deng, T. Li, C. Wang, and M. Yang, "Gated-residual block for semantic segmentation using RGB-D data," *TITS*, vol. 23, no. 8, pp. 11 836–11 844, 2021.

[20] W. Zhou, Y. Cai, and F. Qiang, "Morphology-Guided Network via Knowledge Distillation for RGB-D Mirror Segmentation," *TITS*, pp. 17 382–17 391, 2024.

[21] J. Wolleb, R. Sandkühler, F. Bieder, P. Valmaggia, and P. C. Cattin, "Diffusion models for implicit image segmentation ensembles," in *MIDL*. PMLR, 2022, pp. 1336–1348.

[22] J. Wu, R. Fu, H. Fang, Y. Zhang, Y. Yang, H. Xiong, H. Liu, and Y. Xu, "MedSegDiff: Medical image segmentation with diffusion probabilistic model," in *MIDL*. PMLR, 2024, pp. 1623–1639.

[23] J. Wu, W. Ji, H. Fu, M. Xu, Y. Jin, and Y. Xu, "MedSegDiff-V2: Diffusion-based medical image segmentation with Transformer," in *AAAI*, vol. 38, 2024, pp. 6030–6038.

[24] Z. Xing, L. Wan, H. Fu, G. Yang, and L. Zhu, "Diff-UNet: A diffusion embedded network for volumetric segmentation," *arXiv preprint arXiv:2303.10326*, 2023.

[25] A. Rondinella, F. Guarnera, O. Giudice, A. Ortis, G. Russo, E. Crispino, F. Pappalardo, and S. Battiato, "Enhancing multiple sclerosis lesion segmentation in multimodal MRI scans with diffusion models," in *BIBM*. IEEE, 2023, pp. 3733–3740.

[26] X. Guo, Y. Yang, C. Ye, S. Lu, B. Peng, H. Huang, Y. Xiang, and T. Ma, "Accelerating diffusion models via pre-segmentation diffusion sampling for medical image segmentation," in *ISBI*. IEEE, 2023, pp. 1–5.

[27] Y. Zhao, J. Li, L. Ren, and Z. Chen, "DTAN: Diffusion-based Text Attention Network for medical image segmentation," *CBM*, vol. 168, pp. 1–12, 2024.

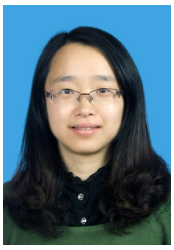
[28] Z. Dong, G. Yuan, Z. Hua, and J. Li, "Diffusion model-based text-guided enhancement network for medical image segmentation," *ESWA*, vol. 249, pp. 1–18, 2024.

[29] Z. Huang, J. Li, N. Mao, G. Yuan, and J. Li, "DBEF-Net: Diffusion-based boundary-enhanced fusion network for medical image segmentation," *ESWA*, pp. 1–13, 2024.

[30] T. Liu, M. Zhang, L. Liu, J. Zhong, S. Wang, Y. Piao, and H. Lu, "CriDiff: Criss-cross injection diffusion framework via generative pre-train for prostate segmentation," in *MICCAI*, 2024, pp. 102–112.

[31] Y. Fu, Y. Li, S. U. Saeed, M. J. Clarkson, and Y. Hu, "Importance of aligning training strategy with evaluation for diffusion models in 3d multiclass segmentation," in *MICCAI*. Springer, 2023, pp. 86–95.

[32] W. Huang and F. Liu, "HiDiffSeg: A hierarchical diffusion model for blood vessel segmentation in retinal fundus images," *ESWA*, vol. 253, pp. 1–13, 2024.
[33] L. Zbinden, L. Doorenbos, T. Pissas, A. T. Huber, R. Sznitman, and P. Márquez-Neila, "Stochastic segmentation with conditional categorical diffusion models," in *ICCV*, 2023, pp. 1119–1129.
[34] Y. Ji, Z. Chen, E. Xie, L. Hong, X. Liu, Z. Liu, T. Lu, Z. Li, and P. Luo, "DDP: Diffusion model for dense visual prediction," in *ICCV*, 2023, pp. 21 741–21 752.
[35] Z. Chen, K. Sun, and X. Lin, "CamoDiffusion: Camouflaged object detection via conditional diffusion models," in *AAAI*, vol. 38, 2024, pp. 1272–1280.
[36] Y. Wen, X. Ma, X. Zhang, and M.-O. Pun, "GCD-DDPM: A generative change detection model based on difference-feature guided DDPM," *TGRS*, pp. 1–16, 2024.
[37] X. Ding, J. Qu, W. Dong, T. Zhang, N. Li, and Y. Yang, "Graph representation learning-guided diffusion model for hyperspectral change detection," *GRSL*, pp. 1–5, 2024.
[38] M. Wang, H. Ding, J. H. Liew, J. Liu, Y. Zhao, and Y. Wei, "SegRefiner: Towards model-agnostic segmentation refinement with discrete diffusion process," *NeurIPs*, vol. 36, pp. 79 761–79 780, 2023.
[39] J. Nam, G. Lee, S. Kim, H. Kim, H. Cho, S. Kim, and S. Kim, "Diffusion model for dense matching," in *ICLR*, 2024, pp. 1–24.
[40] Y. Ye, K. Xu, Y. Huang, R. Yi, and Z. Cai, "DiffusionEdge: Diffusion probabilistic model for crisp edge detection," in *AAAI*, vol. 38, no. 7, 2024, pp. 6675–6683.
[41] B. Ke, A. Obukhov, S. Huang, N. Metzger, R. C. Daudt, and K. Schindler, "Repurposing diffusion-based image generators for monocular depth estimation," in *CVPR*, 2024, pp. 9492–9502.
[42] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.
[43] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *ICLR*, 2021, pp. 1–20.
[44] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *TPAMI*, vol. 34, no. 11, pp. 2274–2282, 2012.
[45] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "CutMix: Regularization strategy to train strong classifiers with localizable features," in *ICCV*, 2019, pp. 6023–6032.
[46] B. Yin, X. Zhang, Z.-Y. Li, L. Liu, M.-M. Cheng, and Q. Hou, "DFormer: Rethinking RGBD representation learning for semantic segmentation," in *ICLR*, 2024, pp. 1–23.
[47] Z. Liu, Y. Wang, Z. Tu, Y. Xiao, and B. Tang, "TriTransNet: RGB-D salient object detection with a triplet transformer embedding network," in *ACM MM*, 2021, pp. 4481–4490.
[48] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *ECCV*, 2018, pp. 3–19.
[49] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *ICCV*, 2017, pp. 4548–4557.
[50] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," in *IJCAI*, 2018, pp. 698–704.
[51] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *CVPR*. IEEE, 2009, pp. 1597–1604.
[52] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *CVPR*. IEEE, 2012, pp. 733–740.
[53] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*. San Diego, CA, USA: Ithaca, 2015, pp. 1–11.
[54] E. Hoogeboom, J. Heek, and T. Salimans, "simple diffusion: End-to-end diffusion for high resolution images," in *ICML*. PMLR, 2023, pp. 13 213–13 232.
[55] S. Chen, C. Ge, Z. Tong, J. Wang, Y. Song, J. Wang, and P. Luo, "AdaptFormer: Adapting vision transformers for scalable visual recognition," *NeurIPs*, vol. 35, pp. 16 664–16 678, 2022.
[56] D. Zhang, K. Song, J. Xu, Y. He, M. Niu, and Y. Yan, "MCnet: Multiple context information segmentation network of no-service rail surface defects," *TIM*, vol. 70, pp. 1–9, 2020.
[57] W. Cui, K. Song, X. Jia, H. Chen, Y. Zhang, Y. Yan, and W. Jiang, "An efficient targeted design for real-time defect detection of surface defects," *OPT LASER ENG*, vol. 178, pp. 1–12, 2024.
[58] S. Zhou, C. Canchila, and W. Song, "Deep learning-based crack segmentation for civil infrastructure: Data types, architectures, and benchmarked performance," *AUTCON*, vol. 146, pp. 1–20, 2023.
[59] F. Guo, Y. Qian, J. Liu, and H. Yu, "Pavement crack detection based on transformer network," *AUTCON*, vol. 145, pp. 1–12, 2023.
[60] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," *NeurIPs*, vol. 34, pp. 12 077–12 090, 2021.
[61] F. Guo, J. Liu, C. Lv, and H. Yu, "A novel transformer-based network with attention mechanism for automatic pavement crack detection," *CBM*, vol. 391, pp. 1–10, 2023.
[62] Z. Liu, Y. Cao, Y. Wang, and W. Wang, "Computer vision-based concrete crack detection using U-net fully convolutional networks," *AUTCON*, vol. 104, pp. 129–139, 2019.
[63] X. Yang, H. Li, Y. Yu, X. Luo, T. Huang, and X. Yang, "Automatic pixel-level crack detection and measurement using fully convolutional network," *CACAIE*, vol. 33, no. 12, pp. 1090–1109, 2018.
[64] S. Zhou and W. Song, "Crack segmentation through deep convolutional neural networks and heterogeneous image fusion," *AUTCON*, vol. 125, pp. 1–16, 2021.

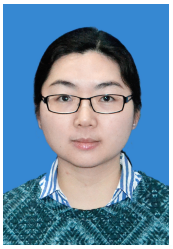


Zhengyi Liu is a professor in School of Computer Science and Technology, Anhui University, China. She received her Ph.D. degree from Anhui University, China in 2007. Her research interests include computer vision and multi-modal large language models.

Junnan Zhou is an M.S. Candidate of Anhui University. He received his B.S. from Jiangsu University of Science and Technology, China in 2024. His research interests include computer vision and deep learning.

Rui Huang is an M.S. Candidate of Anhui University. He received his B.S. from Anhui University, China in 2021. His research interests include computer vision and deep learning.

Xianyong Fang is a professor in School of Computer Science and Technology, Anhui University, China. He received his Ph.D. degree from Zhejiang University, China in 2005. His research interests include computer graphics, virtual reality, computer vision, pattern recognition, multimedia, and human-computer interaction.



Zhengzheng Tu is a professor in School of Computer Science and Technology, Anhui University, China. She received her Ph.D. degree from Anhui University, China in 2015. Her research interests include salient object detection and the other computer vision researches.

Linbo Wang is an associate professor in School of Computer Science and Technology, Anhui University, China. He received his Ph.D. degree from Nanjing University, China in 2014. His research interests include deep learning, computer vision, and image processing.