CrackSegDiff: Diffusion Probability Model-based Multi-modal Crack Segmentation

Xiaoyan Jiang[1*], Licheng Jiang[1*], Anjie Wang[2], Kaiying Zhu[3], Yongbin Gao[1†]

***Abstract*— Integrating grayscale and depth data in road inspection robots could enhance the accuracy, reliability, and comprehensiveness of road condition assessments, leading to improved maintenance strategies and safer infrastructure. However, these data sources are often compromised by significant background noise from the pavement. Recent advancements in Diffusion Probabilistic Models (DPM) have demonstrated remarkable success in image segmentation tasks, showcasing potent denoising capabilities, as evidenced in studies like SegDiff [1]. Despite these advancements, current DPM-based segmentors do not fully capitalize on the potential of original image data. In this paper, we propose a novel DPM-based approach for crack segmentation, named CrackSegDiff, which uniquely fuses grayscale and range/depth images. This method enhances the reverse diffusion process by intensifying the interaction between local feature extraction via DPM and global feature extraction. Unlike traditional methods that utilize Transformers for global features, our approach employs Vm-unet [2] to efficiently capture long-range information of the original data. The integration of features is further refined through two innovative modules: the Channel Fusion Module (CFM) and the Shallow Feature Compensation Module (SFCM). Our experimental evaluation on the three-class crack image segmentation tasks within the FIND dataset demonstrates that CrackSegDiff outperforms state-of-the-art methods, particularly excelling in the detection of shallow cracks. Code is available at https://github.com/skyvisionX/CrackSegDiff.**

I. Introduction

Nowadays, road inspection robots equipping multiple sensors are adopted worldwidely for road structure health monitoring and condition assessment. Among the defects that affect the road health condition, cracks are the most common but challenging type to be detected. Crack segmentation involves identifying cracks on a pavement image at the pixel level, providing accurate shapes and locations of cracks as feedback to the robots. Most machine vision-based crack segmentation algorithms are based on neural network models, for instance, convolutional neural networks (CNN)-based CrackNet-V [3], generative adversarial networks (GAN)-based CrackGAN [4], and SCDeepLab [5] combining CNN and Transformer [6]. The presence of background clutters and the varied appearances of cracks in pavement images pose significant challenges for accurate crack segmentation in practical applications. This variability often leads to misidentification, where cracks are mistakenly recognized as scratches, water streaks, or tar lines. As Fig. 1 shows, depth or range data could provide more information to enhance reliability and yield more consistent results across different environments.

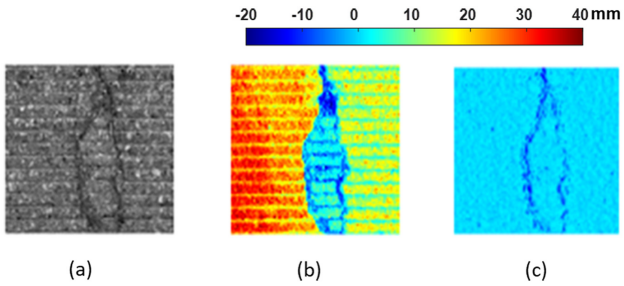

Fig. 1. Example images captured from roadway surface [7]: (a) raw intensity image; (b) raw range image; and (c) filtered range image.

Recently, Diffusion Probabilistic Models have gained much attention and great success in generative tasks, without absolute ground truth. Sequentially, multiple DPM-based segmentors, such as SegDiff [1] and MedSegDiff [8], are introduced to image segmentation using the powerful denoising ability of DPM conditioned on input image. Intuitively, DPM-based segmentor is suitable for the task at hand. But, the reverse diffusion process of the diffusion model in SegDiff has not effectively utilized the information from the original image to guide the denoising process. Noteworthy, due to noise and background variations in pavement images, DPM's U-Net backbone losts crack structural information as the network depth increases during training. This makes DPM-based segmentors insensitive to contextual information, which is crucial for crack segmentation. Hence, Transformer is integrated as the global features to supplement the local features extracted by DPM [9]–[11].

However, directly implementing this approach led to poor performance [8]. One issue is the incompatibility between the abstract conditional features of the Transformer and the features of the diffusion backbone. The Transformer learns deep semantic features from the original image, while the diffusion backbone extracts features from the original image masked with Gaussian noise, making feature fusion more challenging. Additionally, the Transformer's computational complexity, dynamism, and globality make it more sensitive than CNN. Recently, state-space models (SSM) represented by Mamba [12] not only establish long-range dependencies but also maintain linear computational complexity. Vm-unet [2] is a U-shaped architecture model using SSM for image segmentation, capable of capturing extensive global

* Equal contribution.
† Corresponding author: gaoyongbin@sues.edu.cn
[1] School of Electronic and Electrical Engineering, Shanghai University of Engineering Science.
[3] School of Electronic and Computer Engineering, Peking University.
[4] SenseTime.

information and compatible with the U-Net backbone of the diffusion model.

In this paper, we introduce an optimized DPM-based framework for crack segmentation, which integrates grayscale imagery and range data. Utilizing SegDiff [1] as the backbone and a feature enhancement module composed of Vm-unet [2], this model effectively captures long-range dependencies and integrates global features into the U-Net architecture, addressing the lack of global context in SegDiff.

To overcome the challenge of feature compatibility and enhance performance, we have developed the Channel Fusion Module (CFM). This module synergizes multi-scale global and local features from both grayscale and depth images, harnessing their complementary strengths. Grayscale images provide texture details, while depth images offer structural insights, and CFM facilitates their integration through spatial and channel fusion, maximizing the utility of both data types.

Furthermore, detecting shallow cracks remains a challenge due to the loss of low-level features in deeper network layers, exacerbated by noise and background variations in pavement images. To address this, we propose the Shallow Feature Compensation Module (SFCM), which is designed to preserve these critical low-level features, thereby enhancing the segmentation of shallow cracks.

Contributions are summarized as follows: 1) We pioneer the application of a diffusion model for crack segmentation, employing the Feature Enhancement Model (FEM) with Vm-unet for robust global feature integration and enhanced through mixed-loss supervision. 2) We introduce the Shallow Feature Compensation Module, which preserves structural features of cracks by enriching high-level features with multi-scale low-level details, thereby enhancing segmentation accuracy and reducing noise interference. 3) We develop the Channel Fusion Module, designed to seamlessly integrate and optimize the synergy between multi-scale global and local features extracted from both grayscale and depth images. 4) Our CrackSegDiff framework sets a new benchmark for state-of-the-art performance on the FIND dataset across various modalities, including grayscale images, depth images, and their fusion, demonstrating superior crack image segmentation capabilities.

II. Related Work

A. Deep Learning-based Crack Segmentation

Traditionally, crack segmentation relies on CNN to predict the classification label for each pixel. While CNN used in studies, such as [13], [14], and [15], provide reliable baseline performance, they inherently struggle to capture long-range dependencies. This limitation often leads to discontinuous crack detection and false segmentations in complex environments. To address these issues, recent works like [16], [17], and [18] offer more robust solutions by leveraging both local feature and global context modeling. They combine CNN with Transformers, introducing specialized modules to better balance the fusion of local and global features, mitigating the drawbacks of purely CNN-based models.

GAN have also been explored in crack segmentation, offering a way to generate finer details, such as crack boundaries. Studies like [19] and [20] demonstrate the potential of GAN to handle noise and complex backgrounds effectively. However, GAN are often hindered by the implicit nature of their learning process, leading to challenges like unstable training, mode collapse, and artifacts. Research such as [21] and [22] addresses these limitations by employing techniques like penalty terms and joint loss functions to improve performance.

B. DPM in Visual Domain

Recent studies have shown that representations learned by DPM also capture high-level semantic information. Feature maps extracted in the later stages of the reverse diffusion process contain rich representations and are highly effective for segmentation tasks [23]. In the field of medical image segmentation, DPM has achieved new state-of-the-art (SOTA) results on several benchmark datasets [24]. DPM, based on probabilistic modeling, generates images progressively without the need for an adversarial discriminator, addressing some limitations of GAN and gaining popularity in various applications. Studies such as [25], [26], and [27] have applied DPM to medical segmentation. However, the gradual denoising process can lead to the loss of fine details and challenges in maintaining global consistency.

Hence, effectively utilizing useful information from the original image to guide the reverse diffusion process has become a critical challenge. [1] and [9] incorporate original image features to preserve details. However, they do not fully utilize global features or address the complexities of feature fusion, particularly in challenging tasks like crack segmentation. Meanwhile, [28] and [8] incorporate frequency domain guidance, which adds computational complexity and lacks generalization ability across different image types, especially heterogeneous data like fused grayscale and depth images.

C. Grayscale and Depth Image Fusion in Defect Detection

Fusing grayscale and depth images for defect detection presents several challenges, primarily due to the complementary yet distinct nature of the data. Grayscale images are sensitive to lighting and contrast variations, while depth images capture structural information but are often affected by significant background noise from pavements. The key challenge lies in effectively integrating these data sources, minimizing noise, and maximizing the complementary information each image provides. Studies such as [29], [30], [31] have fused grayscale and depth images, leveraging cross-domain feature correlations to achieve more comprehensive defect detection. Despite these advancements, challenges remain in reducing noise and maintaining consistency across the fused data. Diffusion models offer promising potential in this area, as their progressive generation process enables better integration of multi-source information.

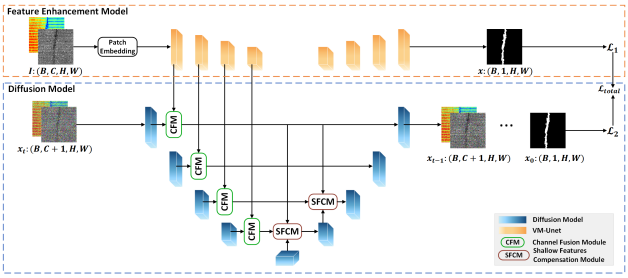

Fig. 2. The overall architecture of CrackSegDiff.

III. The Proposed Method

As shown in Fig. 2, our framework builds upon SegDiff (blue dashed region) based on DPM (III-A) and incorporates Feature Enhancement Model (FEM) to address the challenges in the crack segmentation field. The two modules are highly integrated by a Channel Fusion Module (CFM) in III-C and a Shallow Feature Compensation Module (SFCM) in III-D.

The interaction between the FEM and the diffusion model is essential for ensuring the effectiveness of the incorporated global features. This crucial interaction is mediated by the loss function, which is designed to optimize both the integration of features and the diffusion process. As a result, this strategic control significantly enhances the accuracy of crack segmentation through improved feature augmentation.

A. Diffusion Process of CrackSegDiff

DPM is a generative model parameterized by a Markov chain, consisting of a diffusion process of noises and a reverse diffusion process of reconstruction of original data. The forward process defines the conditional distribution for each step of the reverse diffusion, where each step of the forward process guides the reverse diffusion. Specifically, it is used to define the conditional probability distribution of each step of the reverse diffusion, thereby guiding the model on how to progressively denoise and restore the original data.

Diffusion Process. Given an initial crack data distribution $x_0 \sim q(x_0)$, Gaussian noise $\epsilon \sim N(0,1)$ can be continuously added to the crack segmentation mask x_t at timestamp t :

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \ \bar{\alpha}_t = \prod_{s=0}^{t} \alpha_s \ , \tag{1}$$

where $\alpha_t = 1-\beta_t$. Equ.1 indicates that the standard deviation of x_t is determined by a fixed value β_t. The state prediction is modelled as a Markov chain. As t increases, the final data distribution x_T becomes an isotropic Gaussian distribution:

$$q(x_t|x_{t-1}) := N\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t \mathbf{I}\right), \beta_t \in (0,1) \ , \tag{2}$$

where $\mathbf{I}$ is the identity matrix. The derivation of $q(x_t)$ at any moment can also be entirely based on x_0 and β_t without iteration:

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

Reverse Diffusion Process. This process involves recovering the original crack image from Gaussian noise. Assume the reverse diffusion process is also a Gaussian distribution, it is necessary to construct a parameter distribution function $p_\theta(x_{t-1}|x_t)$ for estimation, as it is not feasible to fit the data distribution incrementally. The reverse diffusion process is a Markov chain process:

$$p_\theta(x_{t-1}|x_t) = N\left(x_{t-1}; \mu_\theta(x_t,t), \sum \theta\left(x_t,t\right)\right) \ . \tag{4}$$

where θ represents the parameters of the reverse process. The key to the reverse diffusion process is designing an effective denoising network to predict the unknown x_0, with the known input x_t and the time encoding t.

B. CrackSegDiff Overview

We model crack image segmentation as a discrete data generation task, aiming to estimate a segmentation map rather than noise. The input consists of grayscale and depth images, along with an additional noise channel, and the output is a refined segmentation map. According to the posterior distribution mean $\tilde{\mu}_t(x_t, x_0)$ [1] in the forward diffusion process

$$\begin{aligned}\tilde{\mu}_t(x_t,x_0) &= \frac{\frac{\sqrt{\alpha_t}}{\beta_t}x_t + \frac{\sqrt{\bar{\alpha}_t}}{1-\bar{\alpha}_t}x_0}{\frac{\alpha_t}{\beta_t} + \frac{1}{1-\bar{\alpha}_{t-1}}} \\ &= \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}x_0\end{aligned} \ , \tag{5}$$

since x_0 is unknown to the network, the proposed model predicts the segmentation map x_0 directly instead of the noise.

Multi-scale features allow the model to detect both small, fine cracks and larger structural patterns, ensuring detailed and robust segmentation across varying crack sizes and complex backgrounds. To better introduce crack image features, we extract multi-scale global features $F_g : [\mathbb{R}^{B\times iC\times \frac{H}{2^i}\times\frac{W}{2^i}}]_{i=1}^{5}$ (where i is the scale and B is the batch size) from the crack data through a FEM's encoder of the same size as the diffusion model encoder. Simultaneously, given the crack image data $I : \mathbb{R}^{B\times iC\times H\times W}$. The crack original image data and the noise are concatenated along the channel dimension as x_t, which is input to the encoder of the diffusion model to obtain the multi-scale local features $F_l : [\mathbb{R}^{B\times iC\times \frac{W}{2^i}\times\frac{H}{2^i}}]_{i=1}^{5}$.

Since F_g and F_l contain the same number and size of features, we merge features of the corresponding scales through the feature fusion module (CFM, details in III-C) to obtain fused features. Subsequently, the fused multi-scale features are supplemented with low-level features through the shallow feature compensation module (SFCM, details in III-D) and input into the decoder of the diffusion model.

We calculate x_{t-1} by:

$$\begin{aligned}x_{t-1} = \alpha_t^{-\frac{1}{2}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(concat(\boldsymbol{I}, x_t), \boldsymbol{I}, t)\right) \\ +\mathbb{1}_{[t>1]}\tilde{\beta}_t^{\frac{1}{2}}z, z \sim N(0,\mathbf{I}), \tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t\end{aligned} \tag{6}$$

to obtain the predicted segmentation map $x_0 \in \mathbb{R}^{B\times C\times W\times H}$. When $t > 1$, $\mathbb{1}_{[t>1]} = 1$.

Finally, the following gradient descent is used until the network converges:

$$\nabla_\theta ||\epsilon - \epsilon_\theta(concat(\mathbf{I}, x_t), \mathbf{I}, t)|| \ . \tag{7}$$

C. Channel Fusion Module (CFM)

So far, while the diffusion model and FEM extract rich local and global features, respectively, the network struggles to fully utilize the spatial co-registration of grayscale and depth/range images. To address this, we introduce a channel fusion module to effectively balance and integrate these features, as shown in Fig. 3.

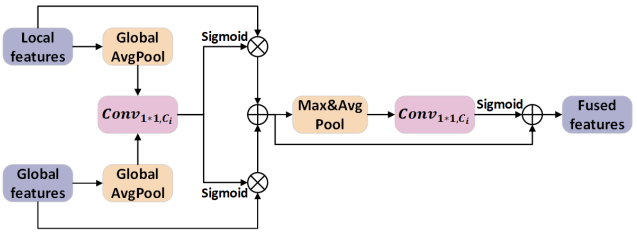

Fig. 3. The overview of CFM.

As previous section, denote the local and global features extracted by the network as F_l and F_g , respectively. First, given that the input channel size of the i-th CFM module is C_i, these features are used for multi-scale feature extraction. The diffusion model extracts local features, while the FEM extracts global features, which are then aggregated through a global pooling layer. This allows the diffusion model to initialize with a rough but static global reference, helping to reduce diffusion variance.

Subsequently, a convolution with a kernel size of 1*1 is applied, producing two sets of weights through two different sigmoid functions. These weights are then multiplied element-wise with the local and global features, respectively, to highlight important features and suppress less important ones, thereby enhancing feature representation capability. The resulting features F_m are then fused through the element-wise addition:

$$\begin{aligned} F_m = F_g * Sigmoid(Conv_{1*1,C_i}(AvgPool(F_g)) + \\ F_l * Sigmoid(Conv_{1*1,C_i}(AvgPool(F_l)) \end{aligned} \ . \tag{8}$$

Next, pooling and convolution operations aggregate effective features from the spatial dimension, maximizing the utilization of the spatial co-registration features of grayscale and depth images, yielding the final fused features $F_m{}'$:

$$\begin{aligned} F_m{}' = F_m * Sigmoid(Conv_{1*1,C_i}(AvgPool(F_m) + \\ MaxPool(F_m)) \end{aligned} \ . \tag{9}$$

D. Shallow Feature Compensation Module (SFCM)

We observed that as network depth increases, some essential low-level features are lost. These features, derived directly from the original image data, are fundamental as they pertain to the physical properties of cracks, encapsulating crucial structural and textural information. The loss of these low-level features poses significant challenges in detecting shallow cracks. Given their high noise levels and minor representation in the overall feature set, incorporating these features into deeper network layers may seem beneficial. However, this integration can also introduce additional noise, complicating the prediction of shallow cracks.

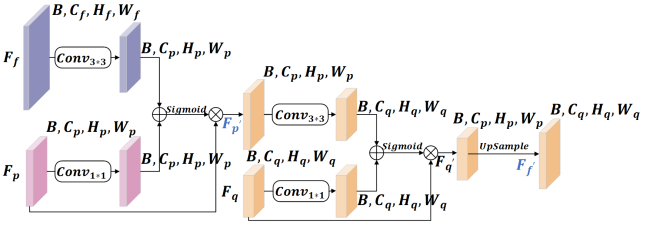

Fig. 4. The overview of SFCM.

To solve this issue, we propose a Shallow Feature Compensation Module to leverage the excellent noise resistance capabilities of diffusion model, as shown in Fig. 4.

$$F_{p'} = F_p * Sigmoid(Conv_{3*3}(F_f) + Conv_{1*1}(F_p)) \ , \tag{10}$$

$$F_{f'} = F_q * Sigmoid(Conv_{3*3}(F_{p'}) + Conv_{1*1}(F_q)) \ . \tag{11}$$

SFCM integrates three distinct feature maps as inputs: F_f from the initial layer following the fusion of the diffusion model and FEM, F_p from the $(n-1)$-th skip connection, and F_q from the n-th layer decoder. This integration is vital for accurately segmenting shallow cracks, which tend to lose detail in deeper layers. F_f enhances both F_p and F_q by adding a substantial quantity of low-level features. This enrichment process culminates in the creation of the intermediate fused feature map $F_{p'}$, and ultimately, the final fused feature map $F_{f'}$.

SFCM enriches and supplements high-level features with multi-scale shallow features while retaining the original features, ensuring that detail-oriented information like shallow cracks does not weaken and disappear in deeper layers of the network. This effectively enhances low-level feature representation. Additionally, supported by the noise-resistant properties of the diffusion model, the module is less susceptible to extra noise in low-level features, thus improving the detection performance of shallow cracks.

E. Loss function

To ensure that SFM and SFCM inject effective features from FEM for the diffusion model, CrackSegDiff is trained using the total loss $\mathcal{L}_{total}$, which combines the classical diffusion noise prediction MSE loss $\mathcal{L}_1$ and the supervised feature supplementation network loss $\mathcal{L}_2$:

$$\mathcal{L}_1 = \mathcal{L}_{mse}(\widehat{x_0}, x_0) \ , \tag{12}$$

$$\mathcal{L}_2 = \mathcal{L}_{dice}(\widehat{x_0}, x_0) + \mathcal{L}_{bce}(\widehat{x_0}, x_0) \ , \tag{13}$$

$$\mathcal{L}_{total} = \alpha\mathcal{L}_1 + \beta\mathcal{L}_2 \ , \tag{14}$$

TABLE I

COMPARISON OF CRACKSEGDIFF WITH STATE-OF-THE-ART GRAYSCALE AND DEPTH FUSED SEGMENTORS ON THE FIND DATASET.

	Raw intensity			Raw range			Fused raw image		
	F1 score	IoU	BF score	F1 score	IoU	BF score	F1 score	IoU	BF score
DenseCrack [3]	*68.2%*	56.5%	-	78.4%	65.3%	-	81.5%	69.7%	-
SegNet-FCN [32]	*75.0%*	63.4%	-	81.1%	68.6%	-	84.0%	72.9%	-
CrackFusionNet [31]	*77.8%*	66.5%	-	82.6%	71.3%	-	86.8%	77.3%	-
Unet-fcn [33]	80.57%	71.25%	84.44%	84.86%	74.69%	87.44%	89.84%	82.53%	91.56%
HRNet-OCR [34]	78.55%	67.73%	85.13%	84.89%	74.18%	89.47%	85.07%	75.55%	90.05%
Crackmer [35]	76.54%	64.92%	81.48%	81.78%	69.72%	84.79%	87.32%	78.25%	89.93%
CT-CrackSeg [36]	*83.55%*	74.39%	*88.61%*	88.51%	80.17%	91.85%	92.75%	87.06%	95.03%
MedSegDiff [28]	83.05%	*74.61%*	88.21%	*90.87%*	*83.70%*	*92.98%*	*95.03%*	*90.77%*	*96.50%*
CrackSegDiff (Ours)	**84.59%**	**77.31%**	**89.23%**	**92.18%**	**86.11%**	**93.71%**	**95.58%**	**91.90%**	**96.63%**

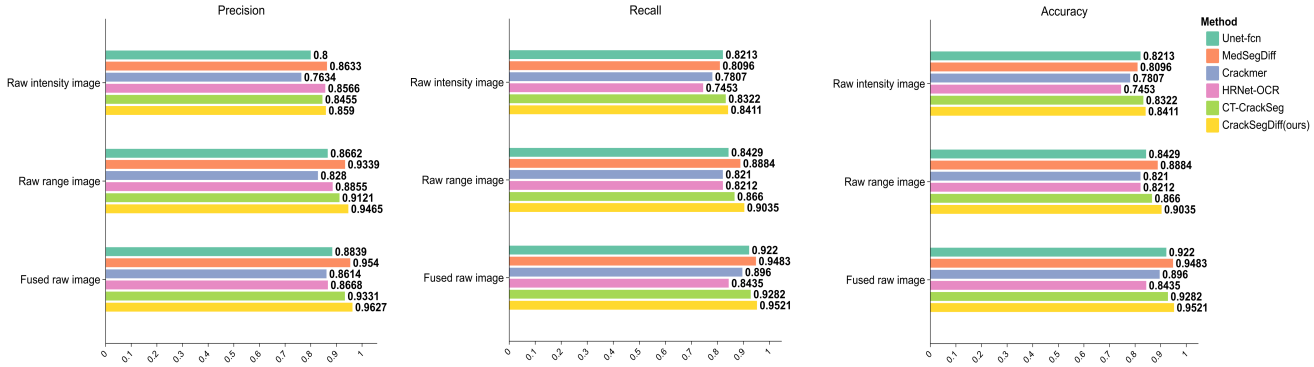

Fig. 5. Comparison charts of CrackSegDiff with state-of-the-art Segmentors on the FIND Dataset using precision, recall, and accuracy.

where α and β are set empirically to 1 and 10, respectively. These values balance the noise prediction loss and supervised feature extraction, ensuring effective feature learning while mitigating noise influence in crack segmentation tasks.

IV. EXPERIMENTS

A. Dataset

We conduct experiments on the FIND dataset [37], which is *currently the only publicly available dataset* to evaluate image fusion-based crack segmentors. Moreover, since FIND captures data from multiple bridge decks and roads under real-world conditions, it is suitable to assess the models' robustness and generalization ability. For one bridge deck or roadway region, FIND provides four data types, that is, grayscale images, range images, filtered range images, and fusion images, which are spatially registered and channel concatenated grayscale and range images. Each data type consists of 2,500 image patches with 256x256 pixels resolution and their corresponding pixel-level ground truth labels.

To verify the noise resistance capability of the proposed model, we did not apply any data pre-processing or augmentation to the original image data. Also, filtered range images in FIND are not used. We randomly divided FIND into a training set and a testing set, containing 2,000 and 500 images, respectively.

B. Experimental Setup

All experiments were conducted in PyTorch and trained and tested using a single NVIDIA A100 GPU. The initial learning rate of the network was set to 1×10^{-4}. The AdamW optimizer with a batch size of 8 was used to search for the optimal segmentation results over 200 epochs. For our model, we employed 1,000 diffusion steps, training in an end-to-end manner. For other methods, 500 images in the training set were used as the validation set. During training, predictions on the validation set were made every 2 epochs to avoid overfitting. In the final testing phase, all models were run once to obtain the final segmentation results.

Segmentation performance was evaluated using the F1-Score [38], IoU [39], and BF-score [40] metrics.

C. Experimental Results

We compare our method with the most advanced segmentation methods presented in the recent FIND dataset review [7], including Unet-fcn [33], a prominent deep learning model for automatic pavement crack segmentation [41], and HRNet-OCR [34], a well-regarded segmentation method in the field. Additionally, we compared our model with other notable methods, including CNN and Transformer combined methods like Crackmer [35] and CT-CrackSeg [36], and diffusion model-based MedSegDiff [28].

Quantitative evaluation. As shown in Table 1, CrackSegDiff ranks the first for all F1-Score, IoU, and BF-Score metrics across all three types of image data in the FIND dataset. Since the boundary distance thresholds of BF-score used in DenseCrack [3], SegNet-FCN [32], and CrackFusionNet [31] are unknown, for fair comparison, we do not compare their BF-scores.

TABLE II

ABLATION STUDY OF MODULES IN CRACKSEGDIFF ON THE FIND DATASET.

	Raw intensity			Raw range			Fused raw image		
	F1 score	IoU	BF score	F1 score	IoU	BF score	F1 score	IoU	BF score
SegDiff-*baseline*	81.95%	71.69%	85.03%	87.46%	78.89%	89.86%	91.94%	84.72%	93.17%
+ SFCM	84.19%	76.93%	88.77%	91.28%	84.78%	93.26%	95.30%	91.39%	96.54%
+ SFM (Proposed)	**84.59%**	**77.31%**	**89.23%**	**92.18%**	**86.11%**	**93.71%**	**95.58%**	**91.90%**	**96.63%**

Our method achieves more effective feature fusion than MedSegDiff. By incorporating global features and supplementing low-level details, it enhances crack structure and fine details, resulting in better performance. As shown in Fig. 5, fused methods obtain better results than single source-based methods. Our model outperforms other methods on all three metrics, that is, precision, recall, and accuracy.

Qualitative evaluation. As the upper part of Fig. 6 shows, Unet-fcn and HRNet-OCR suffer from severe noise interference, low-contrast or blurry regions, leading to incorrect predictions. Crackmer, MedSegDiff, and CT-CrackSeg fail to locate the correct crack positions. However, CrackSegDiff is not affected by noise interference and thus generates segmentation maps with precise details. Furthermore, HRNet-OCR and Crackmer perform poorly when using fused images, failing to effectively utilize the complementary sources.

As the lower sample in Fig. 6 shows, Unet-fcn and Crackmer are easily disturbed by shadows in grayscale images, while Unet-fcn, Crackmer, and MedSegDiff are affected by road grooves in range images. In contrast, our CrackSegDiff demonstrates exceptional noise resistance across all three types of image data.

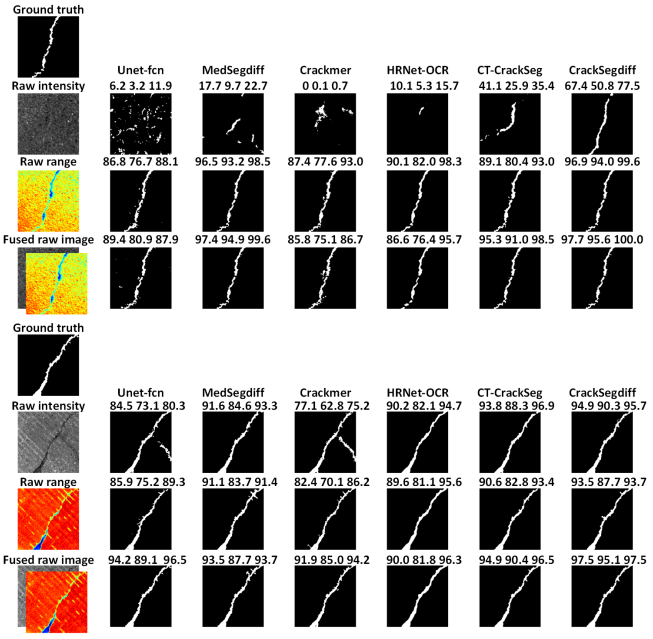

Fig. 6. Qualitative comparison of CrackSegDiff with state-of-the-art segmentation methods. From left to right, the metrics used are F1-Score, IoU, and BF-Score.

D. Ablation Study

As shown in Table 2, the evaluations (F1-Score, IoU, and Bf-Score) are improved progressively as SFCM, and CFM modules are incorporated into the diffusion model. This demonstrates that the SFCM and CFM can further enhance the crack segmentation accuracy of the diffusion model.

Analysis of shallow cracks. The SFCM is specifically designed to address the challenge of recognizing shallow cracks, which is often difficult due to their limited context and subtle features. Although occurrences of such cracks are relatively infrequent in the test set, the impact of the SFCM module on overall performance metrics is not substantial, due to the small number of instances observed.

Despite the limited presence of shallow cracks in the test set, Figure 6 vividly shows the significant impact of the SFCM module. This improvement in a specific scenario underscores the module's effectiveness in enhancing shallow crack detection, suggesting that its benefits, while evident, may not be fully reflected in the overall aggregated performance metrics.

Model illusion. Pavement cracks are often mistaken for plate joints, stains, shadows, and light reflections, which can lead to inaccuracies in model predictions. However, our model addresses this issue by employing fused images that combine spatially registered grayscale and depth data. This integration, alongside the model's step-by-step denoising process, ensures alignment with the true data distribution, preserves realistic details, and minimizes the risk of generating artifacts or false hallucinations.

V. CONCLUSION

In this paper, we introduced a novel framework for crack image segmentation, CrackSegDiff, which leverages a DPM integrated with fused grayscale and range images. To overcome the limitations of traditional DPM-based methods—specifically, their failure to capture global features and difficulties in fusing heterogeneous multi-source data—we developed two pivotal modules: CFM and SFCM. The CFM effectively integrates spatial registration features from both grayscale and range images, enhancing data coherence, while the SFCM specifically targets the improvement of shallow crack segmentation by restoring essential low-level features often lost in standard processing. Experimental results proves the enhanced capability of CrackSegDiff to adeptly manage complex crack segmentation tasks, demonstrating significant robustness against noise and background variations. We hope this paper establishes a new benchmark in crack segmentation and also a new diagram for general DPM-based multimodal image segmentation.

REFERENCES

[1] T. Amit, T. Shaharbany, E. Nachmani, and L. Wolf, "Segdiff: Image segmentation with diffusion probabilistic models," *arXiv preprint arXiv:2112.00390*, 2021.

[2] J. Ruan and S. Xiang, "Vm-unet: Vision mamba unet for medical image segmentation," *arXiv preprint arXiv:2402.02491*, 2024.

[3] Q. Mei and M. Gül, "Multi-level feature fusion in densely connected deep-learning architecture and depth-first search for crack segmentation on images collected with smartphones," *Structural Health Monitoring*, vol. 19, no. 6, pp. 1726–1744, 2020.

[4] K. Zhang, Y. Zhang, and H.-D. Cheng, "Crackgan: Pavement crack detection using partially accurate ground truths based on generative adversarial learning," *IEEE Transactions on Intelligent Transportation Systems (TITS)*, vol. 22, no. 2, pp. 1306–1319, 2020.

[5] Z. Zhou, J. Zhang, and C. Gong, "Hybrid semantic segmentation for tunnel lining cracks based on swin transformer and convolutional neural network," *Computer-Aided Civil and Infrastructure Engineering*, vol. 38, no. 17, pp. 2491–2510, 2023.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[7] S. Zhou, C. Canchila, and W. Song, "Deep learning-based crack segmentation for civil infrastructure: Data types, architectures, and benchmarked performance," *Automation in Construction*, vol. 146, p. 104678, 2023.

[8] J. Wu, W. Ji, H. Fu, M. Xu, Y. Jin, and Y. Xu, "Medsegdiff-v2: Diffusion-based medical image segmentation with transformer," in *AAAI Conference on Artificial Intelligence (AAAI)*, vol. 38, no. 6, 2024, pp. 6030–6038.

[9] G. J. Chowdary and Z. Yin, "Diffusion transformer u-net for medical image segmentation," in *International conference on medical image computing and computer-assisted intervention (MICCAI)*, 2023, pp. 622–631.

[10] J. Zhu, H. Zhu, Z. Jia, and P. Ma, "Diffswintr: A diffusion model using 3d swin transformer for brain tumor segmentation," *International Journal of Imaging Systems and Technology*, vol. 34, no. 3, p. e23080, 2024.

[11] X. Liu, Y. Zhao, S. Wang, and J. Wei, "Transdiff: medical image segmentation method based on swin transformer with diffusion probabilistic model," *Applied Intelligence*, vol. 54, no. 8, pp. 6543–6557, 2024.

[12] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[13] A. Di Benedetto, M. Fiani, and L. M. Gujski, "U-net-based cnn architecture for road crack segmentation," *Infrastructures*, vol. 8, no. 5, p. 90, 2023.

[14] Y. Liu, J. Yao, X. Lu, R. Xie, and L. Li, "Deepcrack: A deep hierarchical feature learning architecture for crack segmentation," *Neurocomputing*, vol. 338, pp. 139–153, 2019.

[15] Y. Fei, K. C. Wang, A. Zhang, C. Chen, J. Q. Li, Y. Liu, G. Yang, and B. Li, "Pixel-level cracking detection on 3d asphalt pavement images through deep-learning-based cracknet-v," *IEEE Transactions on Intelligent Transportation Systems (TITS)*, vol. 21, no. 1, pp. 273–284, 2019.

[16] C. Xiang, J. Guo, R. Cao, and L. Deng, "A crack-segmentation algorithm fusing transformers and convolutional neural networks for complex detection scenarios," *Automation in Construction*, vol. 152, p. 104894, 2023.

[17] C. Wang, H. Liu, X. An, Z. Gong, and F. Deng, "Swincrack: Pavement crack detection using convolutional swin-transformer network," *Digital Signal Processing*, vol. 145, p. 104297, 2024.

[18] J. Quan, B. Ge, and M. Wang, "Crackvit: a unified cnn-transformer model for pixel-level crack extraction," *Neural Computing and Applications*, vol. 35, no. 15, pp. 10 957–10 973, 2023.

[19] Z. Gao, B. Peng, T. Li, and C. Gou, "Generative adversarial networks for road crack image segmentation," in *International Joint Conference on Neural Networks (IJCNN)*, 2019, pp. 1–8.

[20] A. Sekar and V. Perumal, "Cfc-gan: Forecasting road surface crack using forecasted crack generative adversarial network," *IEEE Transactions on Intelligent Transportation Systems (TITS)*, vol. 23, no. 11, pp. 21 378–21 391, 2022.

[21] L. Tian, Z. Wang, W. Liu, Y. Cheng, F. E. Alsaadi, and X. Liu, "A new gan-based approach to data augmentation and image segmentation for crack detection in thermal imaging tests," *Cognitive Computation*, vol. 13, pp. 1263–1273, 2021.

[22] Z. Pan, S. L. Lau, X. Yang, N. Guo, and X. Wang, "Automatic pavement crack segmentation using a generative adversarial network (gan)-based convolutional neural network," *Results in Engineering*, vol. 19, p. 101267, 2023.

[23] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 45, no. 9, pp. 10 850–10 869, 2023.

[24] A. Kazerouni, E. K. Aghdam, M. Heidari, R. Azad, M. Fayyaz, I. Hacihaliloglu, and D. Merhof, "Diffusion models in medical imaging: A comprehensive survey," *Medical Image Analysis (MIA)*, vol. 88, p. 102846, 2023.

[25] J. Wolleb, F. Bieder, R. Sandkühler, and P. C. Cattin, "Diffusion models for medical anomaly detection," in *International Conference on Medical image computing and computer-assisted intervention (MICCAI)*, 2022, pp. 35–45.

[26] J. Wyatt, A. Leach, S. M. Schmon, and C. G. Willcocks, "Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 650–656.

[27] Y. Zhao, J. Li, L. Ren, and Z. Chen, "Dtan: Diffusion-based text attention network for medical image segmentation," *Computers in Biology and Medicine*, vol. 168, p. 107728, 2024.

[28] 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 *Medical Imaging with Deep Learning (MIDL)*, 2024, pp. 1623–1639.

[29] J. Guan, X. Yang, L. Ding, X. Cheng, V. C. Lee, and C. Jin, "Automated pixel-level pavement distress detection based on stereo vision and deep learning," *Automation in Construction*, vol. 129, p. 103788, 2021.

[30] P. Li, B. Zhou, C. Wang, G. Hu, Y. Yan, R. Guo, and H. Xia, "Cnn-based pavement defects detection using grey and depth images," *Automation in Construction*, vol. 158, p. 105192, 2024.

[31] S. Zhou and W. Song, "Crack segmentation through deep convolutional neural networks and heterogeneous image fusion," *Automation in Construction*, vol. 125, p. 103605, 2021.

[32] T. Chen, Z. Cai, X. Zhao, C. Chen, X. Liang, T. Zou, and P. Wang, "Pavement crack detection and recognition using the architecture of segnet," *Journal of Industrial Information Integration*, vol. 18, p. 100144, 2020.

[33] L. Zhang, J. Shen, and B. Zhu, "A research on an improved unet-based concrete crack detection algorithm," *Structural Health Monitoring*, vol. 20, no. 4, pp. 1864–1879, 2021.

[34] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, vol. 43, no. 10, pp. 3349–3364, 2020.

[35] J. Wang, Z. Zeng, P. K. Sharma, O. Alfarraj, A. Tolba, J. Zhang, and L. Wang, "Dual-path network combining cnn and transformer for pavement crack segmentation," *Automation in Construction*, vol. 158, p. 105217, 2024.

[36] H. Tao, B. Liu, J. Cui, and H. Zhang, "A convolutional-transformer network for crack segmentation with boundary awareness," in *IEEE International Conference on Image Processing (ICIP)*, 2023, pp. 86–90.

[37] W. S. S. Zhou, C. Canchila, "Fused image dataset for convolutional neural network-based crack detection (find)," https://doi.org/10.5281/zenodo.6383044, 2022.

[38] T. Fawcett, "An introduction to roc analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

[39] W. I. D. Mining, *Introduction to data mining*, 2006.

[40] G. Csurka, D. Larlus, F. Perronnin, and F. Meylan, "What is a good evaluation measure for semantic segmentation?" in *British Machine Vision Conference (BMVC)*, vol. 27, no. 2013, 2013, pp. 10–5244.

[41] H. Gong, L. Liu, H. Liang, Y. Zhou, and L. Cong, "A state-of-the-art survey of deep learning models for automated pavement crack segmentation," *International Journal of Transportation Science and Technology*, 2023.