Structure- and Style-Preserving Diffusion for High-Fidelity Underwater Crack Generation

Haoyang Li[1], Yongsheng Ou[1*], Kunmo Li[1], and Mingyang Liu[1]
[1] Dalian University of Technology, Dalian 116024, China

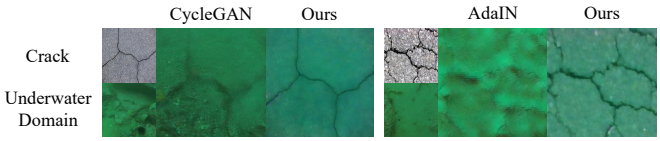
***Abstract*—Bridge structure monitor relies on precise crack detection, yet current deep learning surveillance methods rely heavily on a large number of crack images as dataset, which are difficult to collect in complex and dangerous underwater environments. To alleviate this problem, we propose a training-free diffusion framework that enables efficient and high-quality underwater crack generation. Specifically, we first apply the denoising diffusion implicit models inversion to extract crack and underwater representations. We propose a structure preservation module that leverages global and local information from crack representations to enable the preservation of fine-grained crack details, mitigating structural loss. In parallel, we propose a style integration module to integrate underwater domain information and prevent content leakage issue. To better align the underwater domain style, we design a color adaptation module that dynamically performs adaptive instance normalization, which adjusts the color distribution of the translated result and further optimizes the visual realism of underwater crack images. Extensive experiments show that the proposed framework achieves a significant improvement on generating high fidelity underwater cracks. Code is available at https://github.com/LiHaoyang0517/High-Fidelity-Underwater-Crack-Generation**

***Index Terms*—Style transfer, diffusion models, underwater crack generation**

I. Introduction

Crack detection is a fundamental approach to monitor the health and safety of bridges. With the rapid development of computer vision, deep learning-based perception methods, which possess huge potential in feature extraction, are gradually replacing traditional manual inspection approaches [1]–[3]. Nevertheless, their accurate detection is heavily constrained by the scale and quality of the training data. This limitation is more severe in the underwater scenario where images suffer from color distortion, poor illumination, and low contrast, resulting in scarce and low-quality underwater crack datasets [4]. To alleviate data inadequacy, synthetic data are adopted as an efficient alternative to expensive data collection [5]. Specifically, style transfer has been explored as a promising solution, which is capable of translating terrestrial crack images into underwater bridge cracks [4].

Early style transfer architectures are based on Convolutional Neural Networks (CNNs), in which content and style separation is performed [8]. Later works perform feature statistics for content and style, specified by Adaptive Instance Normalization (AdaIN) [9]. Subsequently, Generative Neural Networks (GANs) [6] introduce adversarial learning to achieve cross-domain mapping, further improving visual quality and fidelity. Despite the success of these conventional methods, they rely on global feature processing and struggle to preserve fine-grained details. This issue is detrimental to structure-sensitive content such as cracks, where geometric continuity is particularly essential [4]. Moreover, they are prone to being affected by the semantics of the style image, causing content leakage problems (see Fig. 1 (a)). Such intrinsic deficiencies hinder their ability to generate high-quality underwater cracks and call for a novel generative paradigm that goes beyond feature-level operations and preserves structure through progressive refinement.



(a) Content Leakage (b) Structure Preservation

Fig. 1. Common problems for current style transfer methods and our improvement. (a) CycleGAN [6] learns the semantics of underwater domains and results in the content leakage problem, while our method is unaffected. (b) AdaIN [7] breaks the continuity and topology of the crack, while our framework preserves the fine-grained details.

Recently, the emergence of diffusion models demonstrates their remarkable capability in high-fidelity image synthesis through step-by-step denoising ability, which enables better preservation of content consistency. Some of the previous works achieve this goal by introducing external guidance. For instance, ControlNet [10] incorporates user-preferred conditions to guide the denoising process, while the plug-and-play mechanism [11]–[13] operates cross-attention maps to achieve prompt-level control. Nevertheless, these methods are highly dependent on external conditions or semantic cues, making them sensitive to style semantics [14]. As a consequence, the structural continuity of the crack will be destroyed. These findings indicate that existing diffusion-based style transfer methods still struggle to simultaneously achieve structural preservation and avoid style semantics bias. Therefore, it is natural to explore whether the step-by-step denoising ability of diffusion models can be exploited to precisely preserve the structure of content, without introducing any external conditions.

To this end, we propose a training-free diffusion pipeline

* Corresponding author: yoo2023@dlut.edu.cn

for underwater crack generation, which aims to preserve fine-grained crack details without the reliance of external conditions. Inspired by the success of the attention mechanism for content and style manipulation [15], [16], we propose a Structure Preservation Module (SPM) that leverages spatial features, *Query*, and *Key* from content images to constrain the denoising process. In parallel, a Style Integration Module (SIM) is designed to incorporate the *Value* component from the underwater domain representations as style information. To further improve visual realism, we design a Color Adaptation Module (CAM) to adjust the color distribution of the generated images, aligning them with the underwater scenes. The proposed architecture enables robust and effective crack detail preservation with minimal influence induced by the underwater style semantics, combines road cracks with underwater scenes, and generates high-quality underwater crack images.

In summary, our contributions are as follows:

- We propose a novel training-free diffusion-based pipeline for translating road cracks into underwater cracks without external conditions.
- We design a Structure Preservation Module (SPM) and a Style Integration Module (SIM) to retain the structural consistency of cracks and mitigate interference caused by underwater semantics.
- We introduce a Color Adaptation Module (CAM) to alter the color distribution of the translated results, enhancing visual realism and fidelity.
- Extensive experiments validate that the proposed method outperforms existing traditional and diffusion-based approaches from both quantitative and qualitative perspectives.

II. Related Work

A. Conventional Style Transfer

In the early stage of style transfer, CNN architectures are predominantly adopted [8], [17]–[20]. [8] pioneers neural style transfer by adopting VGG network [21] to extract features from content and style images. Subsequent works reduce computational complexity and enable efficient style transfer through a single forward pass [22]. AdaIN [9] further enhances quality and flexibility by aligning content and style images via feature statistic operation. GANs [6], [23] are another widely explored framework, known for producing consistent transferred results using adversarial learning. Transformer [24] also plays a significant role in contributing to high quality style transfer. More recently, SANTA [25] proposes shortest path regularization to find a mapping between the source and the target domains. However, these conventional methods require long training time and suffer from instability under insufficient training or inadequate training data.

B. Diffusion-based Style Transfer

Diffusion model has reshaped the research paradigm of style transfer through progressive denoising ability. Early diffusion-based methods focus on content–style disentanglement. StyleDiffusion [26] explicitly extracts content representations while implicitly modeling style information within the diffusion process. InST [27] encodes different styles into text embeddings as conditions for diffusion models. Nevertheless, these methods require additional fine-tuning or training. Training-free methods are proposed to alleviate this problem. DEPM [28] abandons the fine-tuning paradigm by performing patch-based whitening and coloring transformations, while DiffStyle [7] relies on dynamic adjustments to intermediate features in the h-space during generation. Meanwhile, attention-based diffusion frameworks, as flexible means, are widely applied for arbitrary style transfer. Prompt-to-Prompt [11] leverages cross attention maps to modify local semantics and edit images. DiffusionST [15] injects semantic information into the outputs through semantic cues in the self-attention blocks of the denoising U-Net. AnyStyler [29] encodes the content and manipulates the style using a stylized cross attention module. Nonetheless, these diffusion-based methods are far from generating optimal results. One major limitation lies in their semantic-guided nature, where the style-related semantics, introduced through prompts or attention-based injection, tends to distort the original structure of the crack.

C. Underwater Crack Generation

Crack generation has been extensively explored in road and pavement inspection. [31] performs the style transfer between the crack and normal images for data augmentations. CDE-Crack [32] integrates a structure-aware style transfer mechanism to conduct cross-domain adaptation. [33] relies on CNN-based crack translator to synthesize crack image with diverse backgrounds. These studies focus on translating crack images within a single domain; however, extending crack style transfer from terrestrial scenes to underwater environments, which exhibits severe color distortions and low contrast, is highly non-trivial and challenging. [4] utilizes self-constructed above-water and underwater crack datasets and performs unpaired image-to-image translation to synthesize diverse forms of underwater cracks. [5] leverages publicly available crack and underwater datasets and employs CycleGAN [6] for generation. Their works are GAN-based and reveal certain limitations. First, GANs architecture require two unpaired datasets from different domains for training. Without strict constraints between input and output, they can not guarantee the structural fidelity and may lead to inconsistent mappings. Moreover, extensive hyperparameter tuning is typically performed, which highly increases the overall computational cost. On the contrary, our proposed framework adopts a training-free diffusion paradigm, which can efficiently produce high-fidelity underwater crack images.

III. Methods

Given two images, crack image I^c and underwater style image I^s, the goal is to integrate the underwater style from I^s to I^c and generate an underwater crack image I^{cs}. In

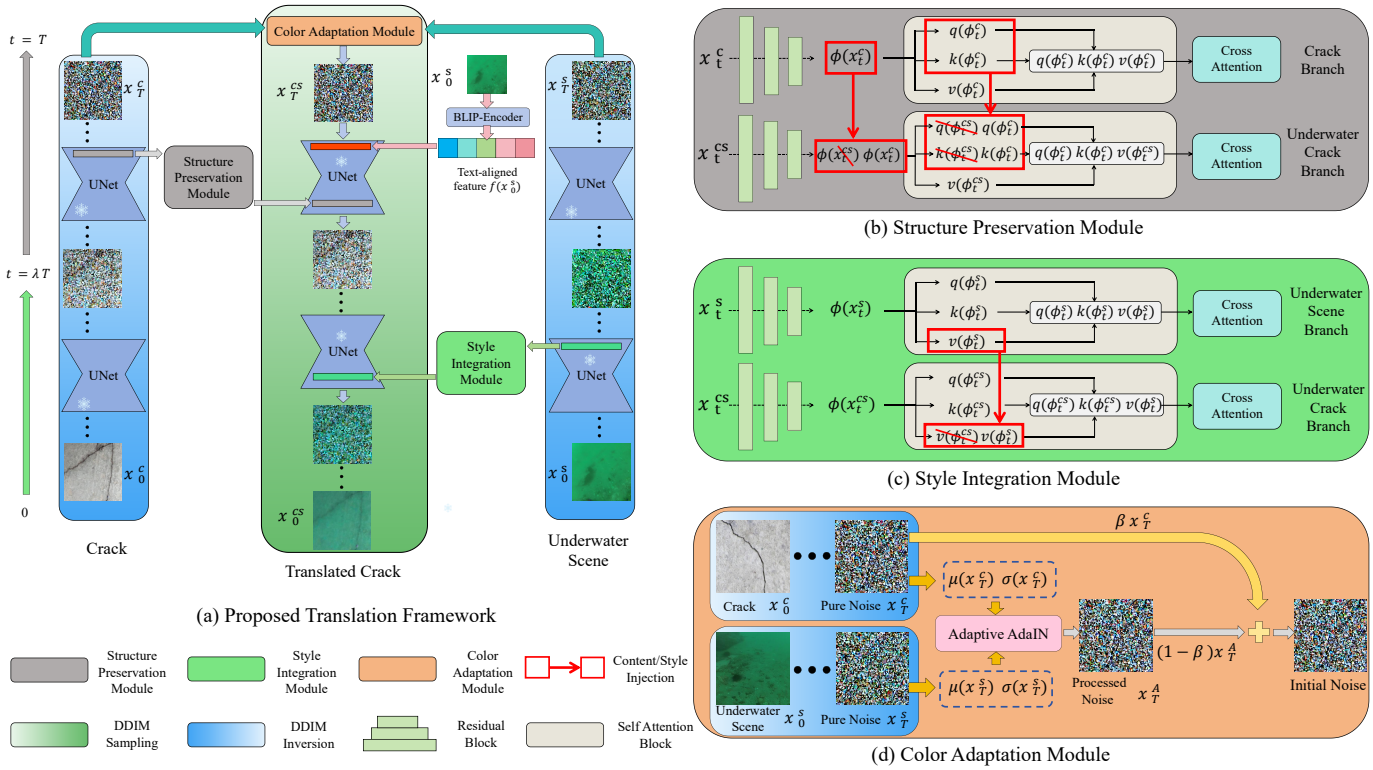


Fig. 2. (a) Proposed Diffusion pipeline for underwater crack translation. (b) Structure Preservation Module. (c) Style Integration Module. (d) Color Adaptation Module. We first use DDIM inversion to convert crack and underwater scene image into noise. At the beginning of sampling, color adaptation module is adopted to rectify the color tone. During the sampling process, BLIP-Encoder [30] will first produce text-embedding of the style image. Then, structure preservation module will inject the residual feature $\phi(x_t^c)$, query $q(\phi_t^c)$ and key $k(\phi_t^c)$ from content image to the translated branch. In terms of the style, style integration module will substitute $v(\phi_t^{cs})$ in the underwater crack branch into $v(\phi_t^s)$ from style image. These two operations help preserve the structure of the crack and inject style information to the target.

this process, the style of I^s and the structural information of I^c should be properly transferred and maintained. Previous learning-based methods require a lot of training time, but achieve a relatively poor performance. Therefore, we are inspired to propose a training-free diffusion framework that extracts content and style representations separately and injects them during the denoising stage. Specifically, the noise latents of crack and underwater images will be effectively produced using Denoising Diffusion Implicit Models (DDIM) inversion. In the denoising stage, the Structure Preservation Module(III-B) and the Style Integration Module(III-C) are used to inject crack and underwater representations based on time steps. The Color Adaptation Module(III-D) is designed to adjust the color distribution of the results and improve their perceptual fidelity. Fig. 2 demonstrates the entire framework.

A. Preliminary

Our method utilizes BLIP-Diffusion [30] and performs diffusion in the latent space. The images I^c and I^s are first encoded into latent representations x_0^c and x_0^s. Then, a forward diffusion is performed in which Gaussian noise is progressively added to the latent representations, producing a noise sequence $x_0^c, x_1^c, ..., x_T^c$. By contrast, in the backward process, the noise is gradually removed, and the generation process is guided by crack and underwater representations. In the following sections, to better elaborate our method, we denote the crack image, the underwater image, and the translated image as x_0^c, x_0^s, and x_0^{cs}, respectively.

B. Structure Preservation Module

In a T-step diffusion process, given an image I, Gaussian noise is progressively added to the image, resulting in intermediate latent x_t at every timestamp and final latent x_T. In the sampling process, a noise-removal network U-Net ε_θ is applied to produce the step by step result x_{t-1}, given by the current latent x_t, current time step t, and condition v. [12] has shown that its residual blocks and self-attention blocks contain spatial information and semantics of the content, which is essential to depict the structural details of the crack. As a result, we propose the structure preservation module and aim to leverage this information. In particular, residual blocks contain high-level representations of the crack (e.g., shape, contour, and boundaries), termed ϕ_t. Followed by the residual blocks, the self-attention blocks will process ϕ_t with a self-attention operation and generate *Query*, *Key*, and *Value*, which are the projections of ϕ_t and contain local information of the crack. As illustrated in Fig. 2 (b), the SPM will replace $\phi(x_t^{cs})$ with $\phi(x_t^c)$ to maintain global crack patterns. *Query* and *Key*

are replaced simultaneously for the preservation of local crack details [16]. The manipulation of the SPM can be expressed as:

$$\phi(x_t^{cs}) = \phi(x_t^c) \tag{1}$$

$$\varphi_t^{cs} = Attention(Q^c, K^c, V) = Softmax(\frac{Q^c K^{cT}}{\sqrt{d}})V \tag{2}$$

in which $\phi(x_t^c)$, Q^c, and K^c are spatial representations, *Query* and *Key* of the crack. Substitutions are performed on the basis of time step and we introduce a hyperparameter λ, termed style strength, to strike a tradeoff between content and style. In a T-step DDIM reverse sampling process, crack representations are injected from T to $\lambda \cdot T$, while the remaining steps, from $\lambda \cdot T$ to 0 are dedicated to the integration of underwater representations. λ does not control the injection of spatial representations in residual blocks because the global structure of the crack should be preserved as much as possible.

C. Style Integration Module

The denoising network ε_θ relies on a condition v (e.g., manual or image-aligned text prompt) to remove noise and produce an outcome. But manual prompts are inflexible and often lead to a failed translation. We apply BLIP-Encoder [30] to produce the text-embedding of the style image, making the generation of v vary from image to image. However, its ability is limited and is not fully compatible with the crack. Hence, we introduce a style integration module to integrate underwater information. As shown in Fig. 2 (c), we extract the *Value* from the underwater branch and inject it to the target branch with operations:

$$\varphi_t^{cs} = Attention(Q, K, V^s) = Softmax(\frac{QK^T}{\sqrt{d}})V^s \tag{3}$$

Similar to the SPM, V^s is the *Value* in self-attention blocks of the underwater representations. The *Value* contains the least spatial information in underwater image which prevents the content leakage problem. Meanwhile, it depicts the visual effect of the underwater image which can be used to align with the crack [12]. The integration of underwater style still relies on λ. By adjusting the hyperparameter, the proposed pipeline can flexibly control the extent to which underwater representations are transferred to the target image.

D. Color Adaptation Module

We observe inconsistent color distributions in the translated results. [16] discovers a similar problem, where style loss occurs in style transfer. Inspired by the work, we design a Color Refinement Module (CAM) (see Fig. 2 (d)), which comprises the Adaptive AdaIN and Adaptive Summation to adjust the color distribution of latent representations at the beginning of sampling.

The principle of Adaptive AdaIN is similar to the original AdaIN, while we introduce a rectification index β between 0 and 1 to adjust the color distribution of the result. To avoid numerical instability when β approaches 0, a small constant $\gamma = 0.001$ is added to the denominator. In equation 4, the Adaptive AdaIN will produce merged noise x_T^A.

$$x_T^A = \frac{x_T^c - \mu(x_T^c)}{\sigma(x_T^c)}\sigma(x_T^s) + \frac{\mu(x_T^s)}{\beta + \gamma} \tag{4}$$

where $\mu(x_T^c)$, $\mu(x_T^s)$ are the mean, $\sigma(x_T^c)$, $\sigma(x_T^s)$ are the standard deviation of the crack and the underwater latent encoding. Moreover, we propose the Adaptive Summation which performs a summation of x_T^A and x_T^c to avoid losing too much fidelity.

$$x_T^{cs} = (1 - \beta) \cdot x_T^A + \beta \cdot x_T^c \tag{5}$$

Similarly, by adjusting β, CAM can dynamically balance two portions, effectively modulating the color distribution of the initial latent.

IV. Experiments

A. Datasets

To verify the performance of the proposed method, we select three publicly available road crack datasets as content while a real-world underwater dataset acts as style to perform underwater crack translation.

Crack500: Crack500 [34] is a concrete road crack segmentation dataset captured by a smart phone, which contains approximately 5k images with a resolution of 640×360.

DeepCrack: DeepCrack [35] is a crack segmentation dataset with 521 images in the size of 448×448. It includes various forms of road cracks, ranging from small, clear cracks to large, fuzzy cracks with diverse background.

Xu Crack Dataset: Constructed by [36], it is designed for crack classification and consists of a total of 6k crack images of size 224×224, which embraces minimal background noise and is particularly suitable for underwater crack generation.

RUIE Dataset: The Real-world Underwater Image Enhancement (RUIE) dataset [37] comprises three subsets, among which the Underwater Image Quality Set (UIQS) and the Underwater Color Cast Set (UCCS) are used in this study. They reflect commonly seen underwater image quality problems (e.g., color cast, visibility degradation, low illumination) and are capable of simulating the poor vision quality of the underwater scene.

B. Implementation Details

Our method uses the pretrained BLIP-Diffusion architecture [30] as the base architecture and manipulates the crack and underwater images in the latent space. In terms of the denoising process, we set both the DDIM steps and the sampling steps as 50. The style strength λ is set to 0.8 to inject more underwater style into the translated result. And the rectification index β is set to 0.5, which is a moderate choice. All experiments are performed on a NVIDIA GeForce RTX 4090 GPU with 24GB of memory.

TABLE I

QUANTITATIVE RESULTS WITH CONVENTIONAL METHODS (1^{th}-4^{th} ROWS) AND DIFFUSION METHODS (5^{th}-8^{th} ROWS). L_C REPRESENTS CONTENT LOSS; S_C REPRESENTS CLIP SCORE; **BOLD** INDICATES THE BEST RESULT; UNDERLINE INDICATES THE SECOND BEST RESULT.

Method	Crack500					DeepCrack					Xu Crack Dataset				
	SSIM ↑	L_C ↓	LPIPS ↓	S_C ↑	FID ↓	SSIM ↑	L_C ↓	LPIPS ↓	S_C ↑	FID ↓	SSIM ↑	L_C ↓	LPIPS ↓	S_C ↑	FID ↓
AdaIN	0.149	1.319	0.856	0.724	381.4	0.371	0.669	0.736	0.733	346.6	0.478	0.380	0.653	0.699	297.9
CycleGAN	0.162	1.318	0.895	0.714	482.4	0.403	0.677	0.718	0.704	353.5	0.520	0.337	0.648	0.719	359.3
StyTR^2	0.131	1.404	0.734	0.604	353.3	0.424	0.682	0.667	0.732	274.2	0.499	0.368	0.690	0.720	314.8
SANTA	0.228	1.388	0.800	0.677	364.3	0.428	0.725	0.653	0.660	333.7	0.528	0.418	0.583	0.670	325.4
DiffStyle	0.126	1.537	0.912	0.685	415.4	0.348	0.880	0.739	0.697	364.1	0.451	0.519	0.714	0.721	381.3
DiffuseIT	0.103	1.623	0.821	0.706	349.7	0.280	0.958	0.703	0.718	282.5	0.347	0.721	0.700	0.717	289.9
InST	0.114	2.746	0.787	0.525	444.2	0.311	1.607	0.703	0.577	378.2	0.407	1.160	0.681	0.593	428.8
InstantStyle	0.078	1.895	0.750	0.717	355.5	0.278	1.397	0.671	0.698	315.5	0.211	1.902	0.870	0.709	496.1
Ours	**0.235**	**1.160**	**0.691**	**0.744**	**231.2**	**0.519**	**0.596**	**0.575**	**0.795**	**171.4**	**0.664**	**0.317**	**0.551**	**0.790**	**209.0**

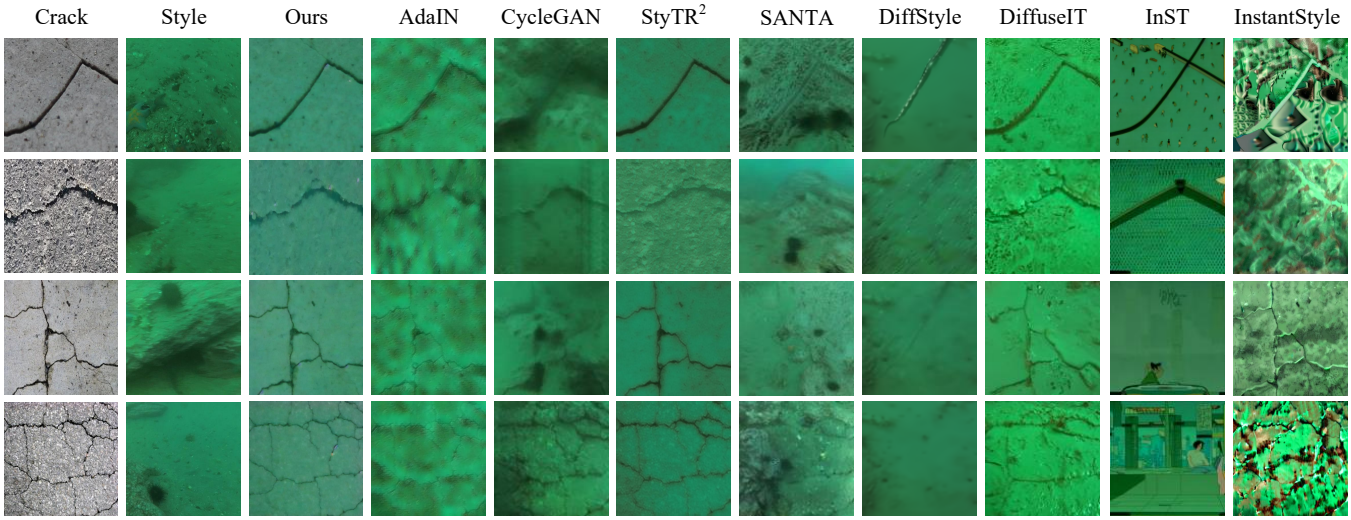


Fig. 3. Qualitative comparison with other methods.

C. Evaluation Metrics

To analyze the performance of all methods for underwater crack translation, we adopt five different evaluation metrics, including SSIM [38], Content Loss [8], LPIPS [39], image-to-image CLIP-Score [40], and Fréchet inception distance (FID) [41].

D. Evaluation Setting and Data Distribution

Since our method as well as some diffusion-based methods do not require training, we can directly apply them on selected datasets. As for methods that require training, we create training and testing sets. Among the crack datasets, Crack500 contains raw training and testing sets. Xu Crack Dataset separates images into three groups, each comprising 2000 images, where the first group acts as a training set and the second one is for testing. We randomly split DeepCrack into two sets with a ratio of 7:3. Regarding RUIE, we use subsets C and D from its UIQS subset, each containing 726 images, as the training set. The testing set comes from the green subset of the UCCS. When evaluating, all translated results are resized to 512×512 for a fair comparison.

E. Comparison with Other Methods

To comprehensively assess our methods, we perform quantitative and qualitative comparisons with multiple style transfer methods. These methods include conventional methods such as AdaIN [9], CycleGAN [6], SANTA [25]; transformer-based method StyTR^2 [24] and SOTA diffusion-based methods, DiffStyle [7], DiffuseIT [42], InST [27] and InstantStyle [43].

1) Quantitative Analysis: As reported in Table I, our proposed method surpasses all other methods among five evaluation metrics across three datasets. In terms of content and structure preservation, the proposed method achieves a SSIM of 0.519 on DeepCrack, significantly outperforms other methods. The best content fidelity can be seen from a relatively low LPIPS, with 0.691, 0.575, and 0.551 on three datasets. Though the result from CycleGAN achieves the second best Content Loss, its FID score is extremely high, revealing a loss of authenticity in the content. Our method achieves a

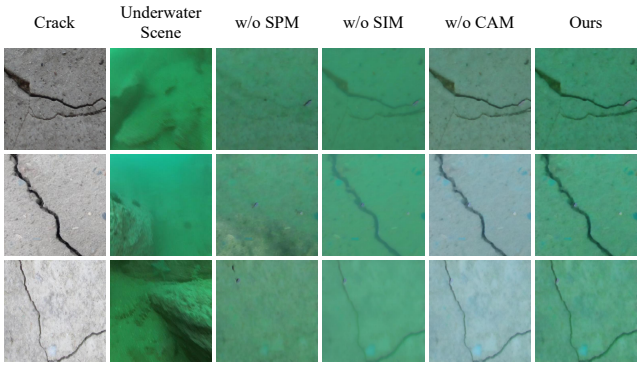


Fig. 4. Abalation study on SPM, SIM and CAM. The translated result shows incomplete structure, strong human artifacts and poor color alignment when SPM, SIM and CAM is missing respectively. Only when three components are combined, the translated images achieve balanced effect.

favorable balance between the preservation of both content and style, yielding the lowest FID scores of 231.2, 171.4, and 209.0 on three datasets. This is further evidenced by the highest CLIP-Score, reaching 0.795 and largely surpassing other methods. Our method consistently demonstrates the best performance by a large margin, even when competing with diffusion-based methods. Some diffusion methods (e.g., InST, InstantStyle) show a significant drop in SSIM and Content Loss, representing severe content distortion and degradation. Their high LPIPS and low CLIP-Score indicate that these methods are unable to properly capture the underwater style.

2) Qualitative Analysis: As shown in Fig. 3, we randomly select 4 pairs of crack and underwater images among all translated results. Traditional methods like CycleGAN and SANTA, which preserve only the contour of the crack, are severely affected by the semantic information of style. StyTR^2 produces comparable results from a visual perspective but suffers from preserving the local details. Also, it reveals strong human artifacts with unnatural background style. Similar issues occur in diffusion-based methods. Method like InstantStyle struggles to interpret the desired scenarios through a description prompt: 'A crack on bridge pier in the turbid water, with poor quality and low illumination' and would lead to pattern collapses. Only DiffuseIT produces relatively good results, yet it generates strong luminance that contradicts the poor visibility and low illumination in the underwater scene. By contrast, our result preserves the structure of the crack while applying the underwater style and overcomes content leakage in existing methods.

F. Ablation Study

To study the effectiveness of SPM, SIM and CAM, we perform the abalation study to observe their influence on the generated images. As depicted in Fig. 4, without SPM, the crack entities are missing from the results, only the underwater style is incorporated into the image. When the SIM is removed, the translated process relies solely on the BLIP-Encoder and shows strong human artifacts, reflected by the unnatural background color in the translated results. Without the CAM, the generated results maintain the crack structure and underwater style, yet their color distributions are different from the underwater scene. With SPM, SIM, and CAM enabled, the translated results reveal preserved fine-grained crack details, harmonious background scenes, and aligned underwater domain color.

V. Conclusion

This paper leverages the potential of the attention mechanism in the diffusion architecture and presents a training-free diffusion style transfer method to solve underwater crack generation, which addresses commonly seen content preservation and content leakage issues occurring in the translated results. This framework systematically integrates crack and underwater representations during the sampling process using the structure preservation module and the style integration module. We further propose a color adaptation module that aligns the color distribution of the results with that of the underwater domain, improving visual realism. Extensive experimental results demonstrate the superiority of the proposed framework for translating underwater cracks over both conventional and diffusion methods. While the proposed framework produces promising results, the quality of the produced underwater crack images can be further improved. Future work will focus on integrating a data preprocessing stage before the style transfer process, particularly by refining the crack content to reduce domain-specific noise (e.g. shadows). This will enhance the robustness of the proposed framework. Another possible direction is related to computational complexity. Although the framework is training-free, the DDIM inversion process introduces additional computational burdens and is not eligible for engineering deployment. Therefore, more efficient diffusion strategies can be investigated.

Acknowledgment

This work is supported by Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (Grant No. JYB2025XDXM117) and Liaoning Province International Science and Technology Cooperation Program (2024JH2/101900001).

References

[1] T. S. Tran, S. D. Nguyen, H. J. Lee, and V. P. Tran, "Advanced crack detection and segmentation on bridge decks using deep learning," *Construction and Building Materials*, vol. 400, p. 132839, 2023.

[2] J. Zhang, S. Qian, and C. Tan, "Automated bridge surface crack detection and segmentation using computer vision-based deep learning model," *Engineering Applications of Artificial Intelligence*, vol. 115, p. 105225, 2022.

[3] D. Hu, T. Yee, and D. Goff, "Automated crack detection and mapping of bridge decks using deep learning and drones," *Journal of Civil Structural Health Monitoring*, vol. 14, no. 3, pp. 729–743, 2024.

[4] B. Huang, F. Kang, X. Li, and S. Zhu, "Underwater dam crack image generation based on unsupervised image-to-image translation," *Automation in Construction*, vol. 163, p. 105430, 2024.

[5] F. Zhang, Y. Gu, L. Yin, J. Song, C. Qiu, Z. Ye, X. Chen, and J. Wu, "Research on the generation and evaluation of bridge defect datasets for underwater environments utilizing cyclegan networks," *Expert Systems with Applications*, vol. 262, p. 125576, 2025.

[6] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[7] J. Jeong, M. Kwon, and Y. Uh, "Training-free content injection using h-space in diffusion models," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 5151–5161.

[8] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2414–2423.

[9] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1501–1510.

[10] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 3836–3847.

[11] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, "Prompt-to-prompt image editing with cross attention control," *arXiv preprint arXiv:2208.01626*, 2022.

[12] N. Tumanyan, M. Geyer, S. Bagon, and T. Dekel, "Plug-and-play diffusion features for text-driven image-to-image translation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 1921–1930.

[13] K. Li, Y. Ou, H. Cai, J. Ning, and M. Qi, "Spatial graph attentional network based place recognition with visual mamba embedding," in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2025, pp. 8842–8849.

[14] R. Xu, W. Xi, X. Wang, Y. Mao, and Z. Cheng, "Stylessp: Sampling startpoint enhancement for training-free diffusion-based method for style transfer," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 18 260–18 269.

[15] Y. Hu, C. Zhuang, and P. Gao, "Diffusest: Unleashing the capability of the diffusion model for style transfer," in *Proceedings of the 6th ACM International Conference on Multimedia in Asia*, 2024, pp. 1–1.

[16] J. Chung, S. Hyun, and J.-P. Heo, "Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 8795–8805.

[17] M. Elad and P. Milanfar, "Style transfer via texture synthesis," *IEEE Transactions on Image Processing*, vol. 26, no. 5, pp. 2338–2351, 2017.

[18] X. Wang, G. Oxholm, D. Zhang, and Y.-F. Wang, "Multimodal transfer: A hierarchical deep convolutional neural network for fast artistic style transfer," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5239–5247.

[19] K. Li, Y. Zhang, J. Ning, X. Zhao, G. Wang, and W. Liu, "Neighborhood consensus guided matching based place recognition with spatial-channel embedding," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 3291–3296.

[20] K. Li, Y. Ou, J. Ning, F. Kong, H. Cai, and H. Li, "Unified depth-guided feature fusion and reranking for hierarchical place recognition," *Sensors*, vol. 25, no. 13, p. 4056, 2025.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[22] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*. Springer, 2016, pp. 694–711.

[23] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[24] Y. Deng, F. Tang, W. Dong, C. Ma, X. Pan, L. Wang, and C. Xu, "Stytr2: Image style transfer with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 326–11 336.

[25] S. Xie, Y. Xu, M. Gong, and K. Zhang, "Unpaired image-to-image translation with shortest path regularization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 10 177–10 187.

[26] Z. Wang, L. Zhao, and W. Xing, "Stylediffusion: Controllable disentangled style transfer via diffusion models," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 7677–7689.

[27] Y. Zhang, N. Huang, F. Tang, H. Huang, C. Ma, W. Dong, and C. Xu, "Inversion-based style transfer with diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 10 146–10 156.

[28] M. Hamazaspyan and S. Navasardyan, "Diffusion-enhanced patchmatch: A framework for arbitrary style transfer with diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 797–805.

[29] W. Liu, "Anystyler: Adding style control in diffusion models to any modality stylization," in *2025 International Joint Conference on Neural Networks (IJCNN)*, 2025, pp. 1–8.

[30] D. Li, J. Li, and S. Hoi, "Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing," *Advances in Neural Information Processing Systems*, vol. 36, pp. 30 146–30 166, 2023.

[31] T. He, J. Zhang, H. Li, and R. Huang, "Research on pavement crack locating based on style transfer and generative adversarial networks," *International Journal of Pavement Engineering*, vol. 27, no. 1, p. 2606911, 2026.

[32] Y. Apedo, H. Tao, W. Gao, C. Xie, and S. Zhao, "Unsupervised domain adaptation for crack segmentation via cross-domain stylization and dual adversarial feature learning," *Journal of Computing in Civil Engineering*, vol. 40, no. 2, p. 04025146, 2026.

[33] J. Zhong, Y. Ma, M. Zhang, R. Xiao, G. Cheng, and B. Huang, "A pavement crack translator for data augmentation and pixel-level detection based on weakly supervised learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 10, pp. 13 350–13 363, 2024.

[34] L. Zhang, F. Yang, Y. D. Zhang, and Y. J. Zhu, "Road crack detection using deep convolutional neural network," in *2016 IEEE international conference on image processing (ICIP)*. IEEE, 2016, pp. 3708–3712.

[35] 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.

[36] H. Xu, X. Su, Y. Wang, H. Cai, K. Cui, and X. Chen, "Automatic bridge crack detection using a convolutional neural network," *Applied Sciences*, vol. 9, no. 14, p. 2867, 2019.

[37] R. Liu, X. Fan, M. Zhu, M. Hou, and Z. Luo, "Real-world underwater enhancement: Challenges, benchmarks, and solutions under natural light," *IEEE transactions on circuits and systems for video technology*, vol. 30, no. 12, pp. 4861–4875, 2020.

[38] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[39] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

[40] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, and Y. Choi, "Clipscore: A reference-free evaluation metric for image captioning," in *Proceedings of the 2021 conference on empirical methods in natural language processing*, 2021, pp. 7514–7528.

[41] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[42] G. Kwon and J. C. Ye, "Diffusion-based image translation using disentangled style and content representation," *arXiv preprint arXiv:2209.15264*, 2022.

[43] H. Wang, M. Spinelli, Q. Wang, X. Bai, Z. Qin, and A. Chen, "Instantstyle: Free lunch towards style-preserving in text-to-image generation," *arXiv preprint arXiv:2404.02733*, 2024.