

Article

FEDM: Feature-Encoding Diffusion Model for Large-Scale SAR Image Inpainting

Junyu Yang, Wenzheng Wang *  and Chenwei Deng 

The School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China;
3420235009@bit.edu.cn (J.Y.); cwdeng@bit.edu.cn (C.D.)

* Correspondence: wang_wenzheng@bit.edu.cn

Highlights

What are the main findings?

- Two core challenges in large-scale SAR image inpainting are identified: inadequate reconstruction of SAR scattering characteristics and insufficient semantic-global coherence.
- A SAR-specific VAE with multi-dimensional joint loss, a cross-attention semantic embedding module, and a dual-condition guided diffusion model with a progressive sliding window strategy are proposed to address these challenges.

What are the implications of the main findings?

- The proposed FEDM enables high-fidelity microwave backscattering reconstruction and semantically consistent large-scale SAR image inpainting.
- It provides an effective solution for large missing region recovery and supports downstream remote sensing applications.

Abstract

With the wide application of generative models in the field of SAR image inpainting, inadequate reconstruction quality of scattering characteristics and insufficient global coherence of semantic logic remain the core challenges of such tasks. To address these issues, this paper proposes a Feature-Encoding Diffusion Model (FEDM). Guided by local valid regions, the proposed model accurately learns the microwave backscattering distribution law of ground features through a SAR-specific Variational Auto-Encoder (SAR-VAE), thus improving the reconstruction accuracy of backscattering statistics. Meanwhile, it integrates semantic embedding and cross-attention mechanism to strengthen the semantic constraints of SAR scenes, ensuring the logical rationality of the ground feature layout. With progressive diffusion generation and sliding window strategy, the model achieves high-quality reconstruction with coherent semantics and consistent global spatial structure for large-scale missing regions. Experiments on public datasets including OSdataset, SEN1-2, SRSDD-v1.0 and MRSSC show that the proposed method achieves excellent performance in terms of scattering characteristic reconstruction quality and globally coherent generation of semantic logic, and realizes high-quality SAR image inpainting.

Keywords: SAR image; image inpainting; diffusion model; feature encoding



Academic Editor: John Trinder

Received: 24 May 2026

Revised: 21 June 2026

Accepted: 26 June 2026

Published: 1 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

High-resolution Synthetic Aperture Radar (SAR) images play an irreplaceable role in key fields such as ocean monitoring, disaster assessment, and national security, thanks to

their all-time and all-weather imaging advantages. However, during the actual acquisition and transmission process, SAR images are prone to local block missing, stripe defects, and invalid regions due to strong scattering interference, equipment failure, data transmission loss, or artificial occlusion. These missing regions damage the integrity and consistency of the image and degrade the performance of downstream tasks including target detection, semantic segmentation, and scene interpretation. Therefore, SAR image inpainting has become a critical task to be solved in the field of remote sensing image processing.

Although current research on SAR image inpainting has made certain progress, as shown in Figure 1, it still faces two core challenges limited by the unique imaging mechanism and data characteristics of SAR.

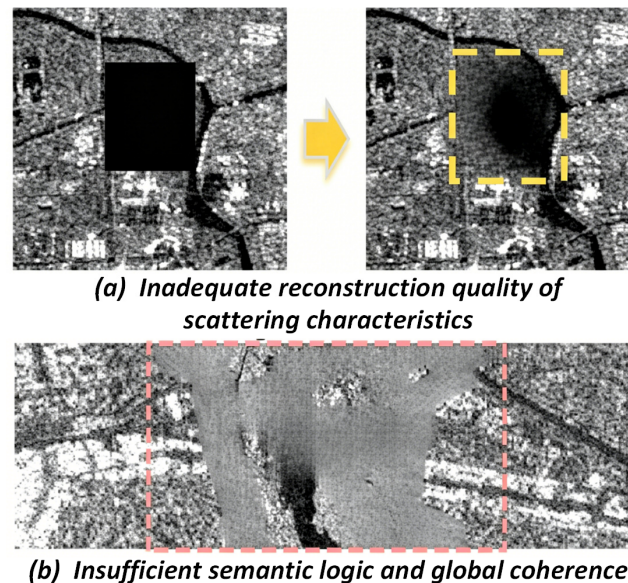


Figure 1. Illustration of the two core challenges in large-scale SAR image inpainting addressed in this work. (a) Inadequate reconstruction of SAR scattering characteristics, with distorted speckle and statistical distribution. (b) Lack of semantic logic and global coherence in large missing regions, causing unnatural scene layouts.

Inadequate reconstruction quality of scattering characteristics: SAR is based on a microwave coherent imaging mechanism. The backscattering characteristics of the ground features show significant complexity and non-stationarity affected by physical properties such as dielectric constant and roughness, and the coherent imaging process introduces strong multiplicative speckle noise. Most existing methods do not design a special modeling mechanism for SAR scattering distribution, and only constrain pixel-level error through simple loss functions. This leads to noise amplification and ground feature structure distortion during the inpainting process. It is not only difficult to reproduce the scattering distribution of real ground features, but also causes structural distortion and scattering distribution offset, resulting in poor consistency of the scattering statistical characteristics (such as the equivalent number of looks and backscattering coefficient) between the inpainted region and the real region.

Insufficient semantic logic and global coherence: SAR images only contain single-channel scattering intensity information and lack rich semantic representation dimensions, making it difficult for existing methods to extract effective scene structure priors from single-channel data. This leads to the lack of accurate semantic constraints in the inpainting process and makes the model prone to problems such as ground feature category conflict and spatial layout disorder. Meanwhile, when dealing with large-scale missing regions, most existing methods adopt simple stitching strategies without establishing an effective

local-global collaborative optimization mechanism. As a result, the local inpainted regions lack global correlation, with obvious stitching seams, artifacts and structural fractures, making it difficult to ensure global spatial continuity.

Further analysis of the existing studies shows that: some methods rely on strongly supervised paired data, but public SAR datasets generally lack large-size and multi-scene inpainting samples, resulting in limited generalization ability of the model and difficulty in adapting to complex real scenes; some methods attempt to introduce semantic constraints, but most directly migrate semantic extraction modules for general images without customized design for SAR data characteristics, leading to limited semantic guidance effect; other methods focus on small-scale missing inpainting, lack targeted optimization for large-scale missing scenes, and find it difficult to balance local detail reconstruction and global coherence.

To solve the above problems, this paper proposes a FEDM based on a diffusion model, which achieves the structure-preserving, scattering-consistent and semantically reasonable SAR image inpainting under weakly supervised conditions. The main contributions of this paper are as follows:

- a. To address the problem of poor fidelity in SAR scattering characteristic reconstruction, we propose a SAR-specific VAE [1] with a multi-dimensional joint loss. We reveal the synergy between KL divergence regularization of the latent space posterior and JS divergence minimization via adversarial loss, which achieves a high-fidelity fitting of the microwave backscattering distributions and improves the reconstruction accuracy of the backscattering statistics.
- b. To tackle the spatial layout disorder caused by the lack of semantic priors in single-channel SAR images, we build a cross-attention semantic embedding module. We find that the attention mechanism can enhance the spatial-semantic adaptive alignment, inject scene priors from the remaining image features, and thus strengthen the semantic consistency in the land-cover categories and spatial layout of the reconstructed features.
- c. To address the global coherence degradation and memory bottleneck in large-scale missing region inpainting, we propose an end-to-end trainable FEDM and design a dual-condition guided diffusion model with a progressive sliding window strategy. We leverage the synergistic denoising mechanism of dual conditions (scattering prior + semantic constraint) in U-Net's multi-scale features, combined with the overlapping consistency constraints of the cyclic sliding window, to achieve coherent reconstruction of the global features and ensure global structural integrity for the large-region inpainting.

The core challenge of large-scale SAR image inpainting lies in simultaneously achieving physically consistent reconstruction of scattering characteristics over wide areas and globally coherent recovery of scene layout, under the constraints of inherent multiplicative speckle noise and single-channel data. Traditional methods rely on handcrafted priors for local interpolation. For example, the Telea algorithm [2] based on the fast marching method is computationally efficient for small defects, but it cannot model long-range spatial dependencies or complex scattering characteristics, resulting in overly smooth outputs with mismatched scattering statistics and discontinuous global structures.

Deep learning methods [3] have become mainstream, yet most still have significant shortcomings in large-scale scenarios. U-Net [4], as a classic encoder–decoder network, can capture local multi-scale context, but its receptive field is limited and it relies solely on pixel-wise loss without dedicated modeling of the SAR backscatter distributions. When handling large missing regions, the physical consistency of scattering and structural integrity degrade severely.

Conditional Generative Adversarial Networks (CGANs) [5] enhance realism through adversarial learning and can balance large-scale structure with local scattering variability. However, their adversarial training is unstable; in a continuous large-area inpainting they are prone to mode collapse, structural distortion, and semantic confusion, failing to maintain global layout consistency over wide SAR scenes [6,7].

Nandakumar et al. introduced SinGAN [8], a single-image generative model, to ISAR ship inpainting. It can reproduce speckle statistics on small, low-resolution patches, but its pyramid-based patch learning paradigm cannot model long-range global spatial correlations in large-scale SAR scenes. It is inherently limited to low-resolution small areas, leading to severe global coherence degradation and semantic mismatch when handling large missing regions.

Cannas et al. applied Deep Image Prior (DIP) [9] to SAR image inpainting. This method maintains good land-cover semantic consistency for small masked areas. However, it lacks dedicated modeling of SAR backscattering mechanisms, relies only on local neighborhood information without global scene priors, and has insufficient fine-grained scattering reconstruction capability for large missing regions. Over wide areas, it is prone to structural distortion, scattering distribution shifts, and semantic incoherence.

Diffusion models [10–12] are gradually being applied to SAR inpainting due to their high-fidelity generation capability. The Latent Diffusion Model (LDM) [13] reduces computational cost via latent mapping and is more suitable for large-scale inputs than pixel-space models. Yet it lacks dedicated encoding for SAR scattering and global semantic layout constraints, and thus cannot accurately fit the backscatter distribution [14]. In large-scale missing inpainting, it suffers from inconsistent scattering reconstruction, fractured spatial structures, and poor global semantic coherence.

However, the above methods have obvious shortcomings in the reconstruction of SAR image scattering characteristics, as well as in the processing of large-scale missing regions and edge detail preservation. In response, this paper focuses on solving the two core problems of accurate reconstruction of scattering characteristics and maintenance of semantic-global coherence based on the feature encoding diffusion framework, and makes up for the deficiencies of the existing methods by integrating SAR-specific feature encoding and semantic enhancement mechanisms.

2. Materials and Methods

2.1. Overall Framework

Aiming at the two core points of SAR image inpainting, namely inadequate reconstruction quality of scattering characteristics and lack of semantic logic and global coherence, this paper constructs an end-to-end SAR image inpainting model based on the stable diffusion framework. As shown in Figure 2, the model consists of four core collaborative modules: SAR-specific Variational Auto-Encoder (SAR-VAE), semantic constraint embedding branch, dual-condition fusion diffusion generation model, and sliding window inpainting strategy.

The training of the model is divided into two stages. In the pre-training stage, SAR-VAE is trained independently with small-format SAR images, and obtains accurate encoding and decoding ability for SAR backscattering statistical representations through adversarial training and joint loss function. After training, the parameters are fixed, and the model is split into an independent encoder and decoder. Then, in the end-to-end joint training stage, the SAR-VAE encoder and decoder, semantic embedding branch, and diffusion generation model are integrated into an integrated generator, which is optimized end-to-end with semantic features as constraints and scattering feature fidelity as the objective. In the inference stage, the trained generator combined with the sliding window

strategy can break through the limitation of hardware video memory and finally achieve high-quality inpainting of missing regions in SAR images.

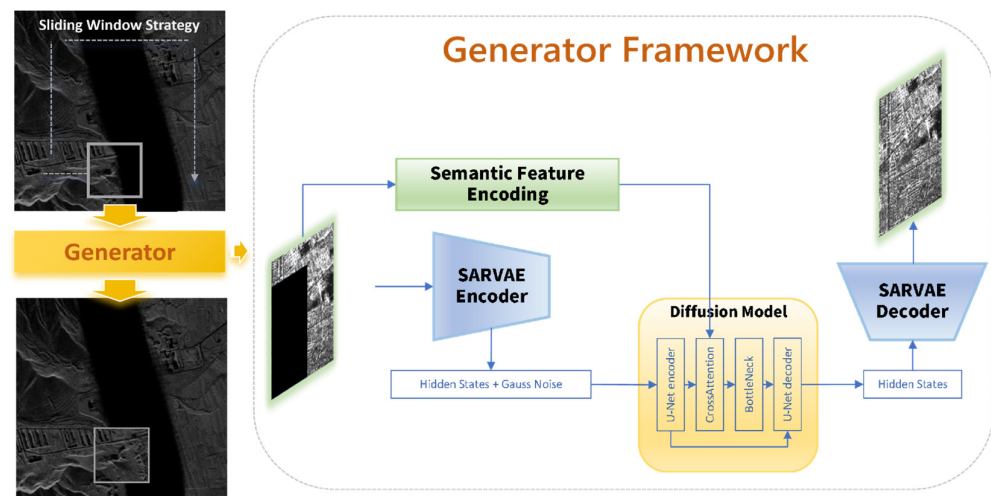


Figure 2. Schematic of the FEDM architecture. This figure illustrates the full inference pipeline with five core components: SAR-VAE encoder, semantic encoder, cross-attention fusion unit, dual-condition diffusion model and SAR-VAE decoder. Masked SAR images are first encoded into latent representations matching ground backscattering statistics by SAR-VAE, while scene semantic features are extracted simultaneously. The two feature streams are adaptively fused through cross-attention and fed into the diffusion model for latent denoising, and the SAR-VAE decoder reconstructs the complete inpainted SAR image.

SAR-VAE fits the feature distribution of real SAR images through a joint loss function and mainly addresses the distribution shift bias; the semantic embedding branch extracts SAR-specific deep semantic features based on VGG16 and injects them into the generation process through a cross-attention mechanism, which specifically solves the semantic logic bias; the diffusion model ensures the feature continuity between locally generated regions and adjacent regions by virtue of multi-scale feature extraction and skip connections of the U-Net structure, and mainly alleviates the coherent generation bias; the sliding window strategy achieves global stitching coherence of ultra-large-format images and further strengthens the mitigation effect on the coherent generation bias.

The four modules work synergistically to ultimately realize the effective suppression of the three types of biases. The following sections will elaborate on the above four components in detail.

2.2. SAR-Specific Variational Auto-Encoder (SAR-VAE)

The core function of SAR-VAE is to realize lossless compression encoding and accurate reconstruction of high-dimensional backscattering statistical representations of SAR images. It avoids computational redundancy and feature distortion of pixel-space diffusion operations through latent space mapping, and accurately fits the scattering feature distribution of real SAR images via a dedicated training paradigm and loss function, thus eliminating the distribution offset between inpainted results and real SAR data. The innovative designs of this module are detailed below.

2.2.1. Exclusive Training Paradigm for SAR Image Inpainting

The standard training paradigm of general Variational Auto-Encoders (VAEs) only takes complete images as input, with an optimization objective limited to full-image reconstruction accuracy. It fails to adapt to the task-level requirements of SAR image inpainting and cannot simultaneously learn scattering feature encoding for valid regions

and feature reconstruction for missing regions, leading to severe feature deviation and performance degradation when directly applied to SAR inpainting scenarios.

To address this limitation, we design an exclusive training paradigm for SAR-VAE to enhance the model's encoding capability for valid regions and scattering feature reconstruction capability for missing regions, with two core designs, as shown in Figure 3:

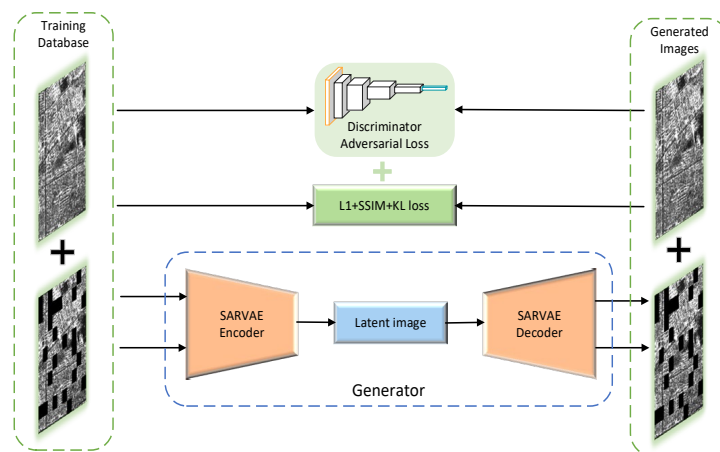


Figure 3. We have constructed the SAR-VAE model to endow it with the capabilities of high-dimensional feature space mapping and feature reconstruction for SAR scene images, providing latent space mapping information for the subsequent training of the diffusion model.

Mask-original image paired training sample construction: We build a paired training sample set tailored for SAR image inpainting. During training, intact high-quality SAR images are randomly selected, and multi-shape masks with 5–60% coverage are generated to simulate common data missing modes in real SAR imaging (e.g., stripe missing, beam occlusion, local information damage), producing corresponding masked SAR images.

The original intact images are used as ground truth to form one-to-one paired training samples. Different from general VAEs trained only on complete images, this strategy drives the model to simultaneously learn accurate scattering feature encoding for valid regions and reasonable reconstruction for missing regions, realizing deep task-level adaptation of the VAE architecture from the training data level.

Dual-module adversarial training architecture: We further construct an adversarial training framework consisting of a SAR-VAE generator and a convolutional discriminator, to solve the scattering statistical mismatch caused by general VAEs relying solely on pixel-level loss. The generator is the SAR-optimized VAE encoding-decoding network, targeting accurate encoding and reconstruction of SAR scattering features. The discriminator adopts a 6-layer convolutional downsampling architecture, which distinguishes real SAR images from SAR-VAE reconstructed outputs and outputs classification confidence via a Sigmoid layer. Through the minimax adversarial game between the two modules, SAR-VAE can accurately fit the backscattering statistical characteristics and inherent speckle distribution of real SAR images while ensuring pixel-level reconstruction accuracy, fundamentally avoiding distribution offset between inpainted results and real SAR data.

2.2.2. Joint Loss Function for Scattering Characteristic Fidelity

A joint loss function consisting of L1 loss, Structural Similarity Index (SSIM) loss, KL divergence loss and adversarial loss is constructed between the generated image and the input image, and the calculation method of each loss term is as follows:

$$Loss_{l1} = \frac{1}{WHC} \sum_{x=1}^W \sum_{y=1}^H \sum_{c=1}^C |I_{in}(x, y, c) - I_{out}(x, y, c)| \quad (1)$$

Equation (1) represents the L1 loss. Here, W and H are the width and height of the image respectively, C is the number of color channels (for grayscale images $C = 1$, for RGB images $C = 3$), I_{in} represents the original SAR image, and I_{outs} represents the generated SAR image.

$$Loss_{KL} = \frac{\mu_p^2 + \sigma_p^2 - 2\log\sigma_p - 1}{2} \quad (2)$$

Equation (2) represents the KL divergence loss, where μ_p and σ_p represent the standard deviation and variance output by the VAE encoder, respectively.

$$Loss_{adv} = -\frac{1}{m} \sum_{i=1}^m (\log(D(z)) + \log(1 - D(G(x)))) \quad (3)$$

Equation (3) represents the generative adversarial loss, where $D(z)$ represents the discriminator's classification score for the original SAR image, and $D(G(x))$ represents the discriminator's classification score for the simulated SAR image generated by the generator.

$$Loss_{ssim} = 1 - \prod_{m=1}^M \left(\frac{2\mu_p\mu_g + c_1}{\mu_p^2 + \mu_g^2 + c_1} \right)^{\beta_m} \left(\frac{2\sigma_p\sigma_g + c_2}{\sigma_p^2 + \sigma_g^2 + c_2} \right)^{\gamma_m} \quad (4)$$

Equation (4) represents the structural similarity loss. M represents different scales. μ_p and μ_g represent the means of the generated SAR image and the original SAR image respectively. σ_p and σ_g represent the standard deviations of the generated SAR image and the original SAR image, respectively. σ_{pg} represents the covariance between the generated SAR image and the original SAR image. β_m and γ_m represent the relative importance between the two terms. c_1 and c_2 are constant terms to prevent the divisor from being zero.

$$Loss_{total} = \lambda_1 Loss_{l1} + \lambda_2 Loss_{KL} + \lambda_3 Loss_{adv} + \lambda_4 Loss_{ssim} \quad (5)$$

The loss calculated between the generated image and the input image is shown in Equation (5), which is obtained by weighted addition of L1, SSIM, KL, and adversarial losses, with weights of λ_1 , λ_2 , λ_3 , and λ_4 , respectively. Through iterative training, the VAE model is enabled to have the encoding and decoding capabilities for masked SAR images and the original images.

2.3. Cross-Attention Semantic Embedding

We design a cross-attention-based semantic embedding branch, which extracts SAR-specific deep semantic information, constructs a fine-grained semantic constraint injection mechanism, and provides full-process semantic guidance for the diffusion model to ensure consistent ground feature categories and logical spatial layout between inpainted and valid regions.

2.3.1. Backbone Network Adaptation

The semantic embedding branch adopts VGG16 as the semantic feature extraction backbone, with exclusive adaptations for the feature differences between single-channel SAR images and optical RGB images:

Input layer adaptation: The number of input channels of VGG16 is modified from 3 to 1 to directly accept single-channel SAR images, avoiding information redundancy and feature distortion caused by single-to-three-channel conversion.

Output layer optimization: The final classification layer of VGG16 is removed, and the output of the 5th convolutional block is used as the semantic extraction result. Inputting a masked SAR image yields a $1 \times N$ semantic feature encoding that integrates high-level semantic information including ground feature category, spatial layout and scattering characteristics, laying a foundation for subsequent semantic constraint injection.

We construct a closed-loop dual-condition diffusion pipeline consisting of two core parts.

2.4.1. Dual-Condition Input Preprocessing and Diffusion Mathematical Formulation

Masked SAR images are fed as input, with binary masks separating missing regions and retaining valid pixels. The pre-trained SAR-VAE encoder maps masked images into latent features Z_{masked} as scattering priors, while the semantic branch extracts semantic vectors as denoising constraints.

The forward diffusion process gradually adds Gaussian noise to clean latent representations:

$$z_t = \sqrt{\alpha_t} z_0 + \sqrt{1 - \alpha_t} \varepsilon, \varepsilon \sim \mathcal{N}(0, I) \quad (7)$$

where z_0 denotes clean latent features encoded from complete SAR images, t is diffusion timestep, and α_t stands for noise schedule coefficients.

Instead of predicting noise like mainstream diffusion models, the U-Net directly reconstructs original clean latent features with the latent reconstruction loss:

$$\mathcal{L}_{latent} = \|z_0 - \text{U-Net}(z_t, F_{sem})\|_1 \quad (8)$$

F_{sem} refers to global semantic features injected into all encoder, decoder and skip layers of U-Net via cross-attention to maintain semantic consistency throughout denoising.

2.4.2. Semantics-Guided Denoising and End-to-End Pipeline

In the reverse denoising phase, semantic features are fixed as persistent conditions. Cross-attention enables deep interaction between semantic embeddings and latent representations across all U-Net blocks and skip connections, imposing global semantic control over the whole generation process. The model is fully optimized via latent L1 loss supervised by ground-truth latent maps of complete SAR data.

During training, paired masked and intact SAR images are input; frozen SAR-VAE extracts latent representations, and the semantic module captures global scene information. Noisy latent variables are fed into U-Net to compute L1 loss for weight update. For inference, only masked images are required; reverse DDIM sampling starts from pure Gaussian noise, and the decoder of SAR-VAE outputs restored SAR results. All 256×256 inputs are encoded into $4 \times 64 \times 64$ latent tensors, with 1000 forward diffusion steps and 50 DDIM sampling steps. We adopt AdamW optimizer with initial learning rate 10^{-4} , weight decay 10^{-5} and batch size 8, training for 200 epochs with 50 warm-up epochs. All configurations are fully documented for reproducibility.

2.5. Sliding-Window Strategy for Large-Format Image Inpainting

To address hardware video memory limits for large-format SAR image input and global coherence degradation in inpainting, we adopt a 50% overlap ratio sliding window strategy. Following the “time-for-space” principle, it decomposes the global inpainting task of large-format images into fixed-size local subtasks.

As shown in Figure 5, this strategy takes the model’s optimal input size N as the window size and $N/2$ as the sliding step, performing global sliding in a fixed order. Each window covers the valid SAR region and the target inpainting mask; local inpainting and result backfilling are completed for each window position, with full image inpainting realized through cyclic iteration.

The video memory consumption of this strategy depends solely on the window size, independent of the original image scale, effectively breaking through hardware limits. The half-window overlapping design eliminates feature fracture at stitching boundaries ensures

global spatial coherence and scattering consistency of inpainted results, and adapts to SAR inpainting tasks of any scale under limited computing resources.

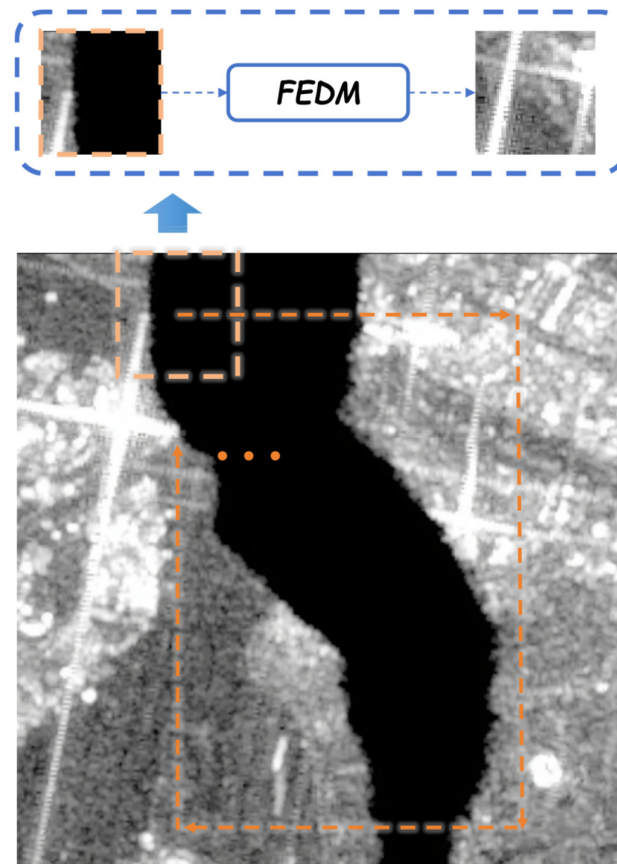


Figure 5. Schematic diagram of sliding window image inpainting strategy: Overall, the cyclic sliding window operation is carried out in the order of “bottom → right → top → left → bottom”. In each sliding window operation, the FEDM is used to generate the area to be completed in the sliding window until the image inpainting is finished.

3. Results

3.1. Datasets

To comprehensively verify the effectiveness and generalization capability of the proposed FEDM for large-format SAR image inpainting tasks, we conduct experiments on three open-source datasets: OS-dataset [15], SEN1-2 [16], and SRSDD-v1.0 [17]. The OS-dataset is a SAR-optical paired dataset with 1–3 m SAR resolution and standard 256×256 patches covering cities, farmlands, mountains and water bodies. The SEN1-2 integrates 10 m-resolution Sentinel-1 C-band SAR and Sentinel-2 optical images with various global land-forms. The SRSDD-v1.0 is a 1–5 m high-resolution SAR ship dataset featuring complex coastal and open-sea clutter. These datasets contain rich ground-object features and serve as reliable benchmarks to evaluate the model’s performance on backscattering statistics reconstruction and semantic consistency. Partial dataset samples are displayed in Figure 6.

Meanwhile, the downstream task validation experiments are carried out based on the Multimodal Remote Sensing Scene Classification (MRSSC) [18] dataset. This dataset is constructed from high-quality Earth observation data acquired by the Tiangong-2 space laboratory, covering seven typical scenarios including urban, farmland, and mountainous areas, with a total of 12,167 optical-SAR matched image pairs.

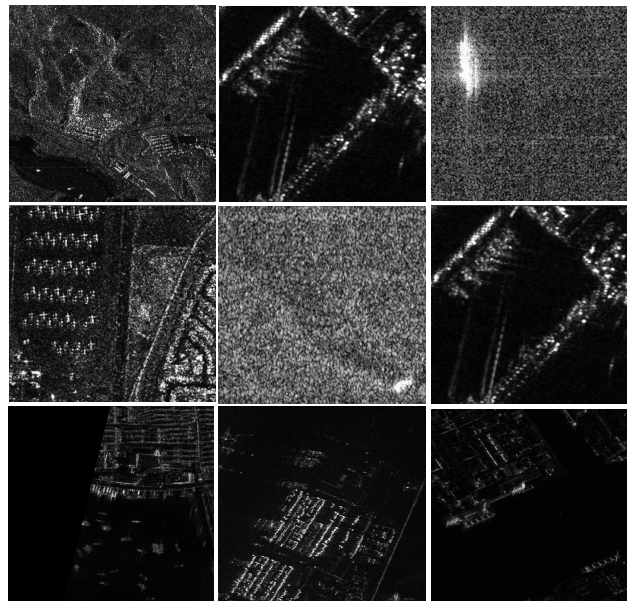


Figure 6. Examples of the SAR image training set are shown in the figure. The images in the dataset contain rich ground object features, which ensure the generalization performance of the model during the training process.

3.2. Evaluation Metrics and Parameter Settings

In this paper, the inpainting quality is evaluated from four dimensions: pixel fidelity, structural similarity, SAR-specific scattering statistical characteristics, and edge preservation capability, using four metrics: Peak Signal-to-Noise Ratio (PSNR) [19], Structural Similarity Index Measure (SSIM) [20], Equivalent Number of Looks (ENL) [21], and Edge Preservation Index (EPI) [22]. All metrics are calculated in accordance with the definitions and implementations in the corresponding references.

The proposed method is implemented based on the PyTorch 1.13.1 (CUDA 11.7) deep learning framework, and all experiments are completed on a single NVIDIA RTX 4090 GPU with 24 GB of video memory.

In the preprocessing stage, all SAR images are uniformly resized to 256×256 pixels, and the scattering intensity values are normalized to the range of $[-1, 1]$. In this paper, missing masks with a coverage rate of 5% to 60% are randomly generated to simulate common missing patterns in real SAR imaging, such as strip missing, beam occlusion, and local information damage.

To obtain optimal hyperparameters for each loss term, we perform grid search with the tuning range set to $[0.1, 1.0]$. The optimal weight combination is selected by comprehensively balancing three quantitative metrics: PSNR, SSIM, and ENL. Considering the imaging characteristics of SAR images, we assign differentiated weights to each component: λ_1 for L1 loss is set to a relatively large value to guarantee pixel-wise reconstruction accuracy; λ_2 of KL divergence loss regularizes the latent posterior distribution and stabilizes the encoding of SAR scattering statistics; λ_3 for adversarial loss is adopted to match the inherent speckle and backscattering distributions of real SAR scenes; λ_4 corresponding to SSIM loss serves as an auxiliary term for edge and structural preservation. The finalized weights are $\lambda_1 = 0.5$, $\lambda_2 = 0.2$, $\lambda_3 = 0.2$, and $\lambda_4 = 0.1$.

3.3. Quantitative Results

Table 1 shows the quantitative comparison results of different methods on the three datasets. Six representative image inpainting methods are chosen for comparison, including traditional algorithms, deep generative models, and recent state-of-the-art diffusion

approaches. They are Telea [2], DIP [9], U-Net [4], CGAN [5], SinGAN [8], and the latent diffusion model (LDM) [13].

Table 1. Quantitative comparison of SAR image inpainting results of different methods.

Method	PSNR	SSIM	ENL	EPI
Telea [2]	18.56	0.281	12.28	0.372
DIP [9]	22.41	0.479	15.66	0.553
U-Net [4]	25.25	0.323	18.40	0.516
CGAN [5]	26.69	0.350	19.77	0.613
SinGAN [8]	25.88	0.431	18.95	0.620
LDM [13]	27.36	0.464	20.08	0.660
Ours	29.62	0.670	24.52	0.714

It can be seen that the proposed FEDM achieves the optimal performance on all evaluation metrics, which fully verifies its advantages in pixel fidelity, consistency of scattering characteristics, structural integrity, and global coherence.

Specifically, compared with LDM, FEDM improves the PSNR by 2.26 dB (from 27.36 to 29.62) and the SSIM by 0.206 (from 0.464 to 0.670), demonstrating significant enhancements in pixel-level accuracy and structural similarity. Meanwhile, the ENL value of FEDM is 24.52, which is substantially higher than the second-best LDM (20.08), proving that the proposed method can better fit the real SAR scattering distribution and effectively suppress speckle noise distortion. Furthermore, FEDM achieves the highest EPI value of 0.714, an improvement of 0.054 over LDM (0.660), demonstrating its strong capability to preserve strong-scattering edges and ship contours.

3.4. Ablation Experiments

To systematically verify the effectiveness of the core designs of the proposed FEDM, this section is divided into three groups of ablation experiments, which respectively verify the value of the semantic embedding branch, the influence of sliding window step size on generation effect, and the role of each term in the joint loss function. All experiments keep the training hyperparameters and test dataset completely consistent with the main experiment, and the core evaluation metrics are PSNR, SSIM, and ENL.

3.4.1. Validation of the Effectiveness of the Semantic Embedding Branch

To verify the role of the cross-attention semantic embedding branch in improving image generation quality, two groups of control experiments are set up in this section: Experimental group 1 is the complete FEDM, and Experimental group 2 is the variant model with the cross-attention semantic embedding branch removed. The quantitative results are shown in Table 2.

Table 2. Validation results of the semantic embedding branch effectiveness.

Model Configuration	PSNR	SSIM	ENL	EPI
FEDM	29.62	0.670	24.52	0.714
Without semantic embedding branch	27.29	0.598	22.63	0.650

After removing the semantic embedding branch, all metrics show a significant decrease: PSNR drops by 2.33 dB (from 29.62 to 27.29), SSIM by 0.072 (from 0.670 to 0.598), ENL by 1.89 (from 24.52 to 22.63), and EPI by 0.064 (from 0.714 to 0.650). Among them, the decrease in SSIM is the most significant, which proves that the branch extracts deep semantic features of single-channel SAR images through the adapted and optimized VGG16,

and injects scene priors into the whole diffusion generation process via the cross-attention mechanism. It can effectively strengthen the constraints on ground feature categories and spatial layout, thus solving the inpainting logic confusion caused by the lack of semantic information in single-channel SAR images.

3.4.2. Analysis of the Influence of Sliding Window Step Size on Generation Effect

This ablation experiment is carried out to explore the influence of sliding step size (window overlap ratio) in the sliding window strategy on the inpainting effect, global spatial coherence, and inference efficiency of large-format SAR images, and to determine the optimal step size setting that balances generation quality and computational efficiency.

The sliding window size is fixed at 256×256 (consistent with the optimal input size of the model), and four groups of different sliding step sizes are set corresponding to different window overlap ratios. The specific configurations are: step size 192 (25% overlap), 128 (50% overlap), 64 (75% overlap), and 32 (87.5% overlap). The experiment is conducted on the large-format SAR test set, with PSNR and SSIM as the evaluation metrics for generation effect, and the average inference time of a single 1024×1024 image is counted at the same time. The quantitative results are shown in Table 3.

Table 3. Performance comparison of different sliding step sizes.

Sliding Step Size	Window Overlap Ratio	PSNR	SSIM	Average Inference Time (s)
192	25%	26.35	0.547	0.78
128	50%	29.62	0.670	1.35
64	75%	29.87	0.663	3.69
32	87.5%	29.94	0.675	12.84

As the step size decreases and the window overlap ratio increases, the PSNR and SSIM of the model show a continuous upward trend. When the step size is 128 (50% overlap), the PSNR and SSIM are close to the upper limit of performance, which are 3.27 dB and 0.063 higher than those of the 25% overlap group, respectively. Meanwhile, the inference time of a single image is only 1.35 s, maintaining high computational efficiency.

As the step size decreases and the window overlap ratio increases, the PSNR and SSIM of the model show an overall upward trend. When the step size is 128 (50% overlap), the PSNR and SSIM are close to the upper performance limit, which is 3.27 dB and 0.123 higher than those of the 25% overlap group, respectively. Meanwhile, the inference time of a single image is only 1.35 s, maintaining high computational efficiency.

When the step size is less than 128 and the overlap ratio exceeds 75%, the improvement of the model's generation quality tends to be flat: the PSNR of the 75% overlap group is only increased by 0.25 dB and SSIM by -0.007 compared with the 50% overlap group, and the PSNR of the 87.5% overlap group is only increased by 0.07 dB and SSIM by 0.012 compared with the 75% overlap group. However, the inference time increases significantly, and the inference time at step size 32 is approximately 9.5 times that at step size 128, resulting in a decrease in engineering practicability.

Considering both generation quality and computational efficiency, this paper finally selects the step size of 128 with 50% overlap ratio as the optimal configuration of the sliding window strategy, which ensures the generation quality while taking into account the inference efficiency.

3.4.3. Validation of the Effectiveness of Each Term in the Joint Loss Function

This ablation experiment is set up to verify the independent contribution of four terms in the proposed joint loss function, namely L1 loss, KL divergence loss, SSIM loss, and adversarial loss, and clarify the influence of each loss term on SAR image inpainting performance.

Taking the complete joint loss function in this paper as the benchmark, control experiments are carried out using the single-variable removal method: only one loss term is removed at a time, the weight ratio of the remaining loss terms is consistent with the main experiment, and the model structure, training strategy, and test dataset are the same for all experiments. The quantitative results are shown in Table 4.

Table 4. Ablation experiment results of each term in the joint loss function.

Configuration	L1 Loss	KL Loss	SSIM Loss	Adversarial Loss	PSNR	SSIM	ENL
Exp. 1	✓	✓	✓	✓	29.62	0.670	24.52
Exp. 2		✓	✓	✓	26.74	0.606	22.17
Exp. 3	✓		✓	✓	28.15	0.583	21.06
Exp. 4	✓	✓		✓	27.84	0.612	23.29
Exp. 5	✓	✓	✓		27.63	0.576	20.30

L1 loss has the most significant impact on image reconstruction accuracy, with a PSNR decrease of 2.88 dB and SSIM decrease of 0.064 after removal, which is the core constraint to ensure pixel-level basic reconstruction accuracy. Adversarial loss has the greatest impact on the physical fidelity of SAR scattering characteristics, with an ENL decrease of 4.22 after removal, which is the key to ensuring the quality of generated images.

The core role of KL divergence loss is concentrated in the SAR scattering feature encoding link, with an ENL decrease of 3.33 after removal. It is the underlying constraint to ensure the rationality of the latent space feature distribution of SAR-VAE, and its importance to physical fidelity is second only to adversarial loss. SSIM loss relatively has the least impact on the core performance of the model, and the decrease in all metrics after removal is lower than that of the other three terms. It mainly plays a supplementary role in optimizing target edges and improving the coherence of structural details.

The experimental results show that the collaborative optimization of the four loss terms achieves the dual goals of physical fidelity of scattering characteristics and semantic coherence of scene structure for SAR images, which verifies the rationality and effectiveness of the multi-dimensional joint loss function design in this paper.

3.4.4. Visualization Results and Analysis

As shown in Figure 7, traditional methods such as Telea and DIP tend to produce over-smoothed inpainting results, losing the detailed texture and scattering characteristics of the scene. The U-Net and CGAN methods generate blurred edges and inconsistent scattering intensity in the inpainted regions. The basic LDM lacks effective semantic constraints, resulting in scene shape distortion and spatial layout confusion.

In contrast, the proposed FEDM can reconstruct the contour and texture details of the scene with high quality. The inpainted region is consistent with the surrounding valid regions in terms of scattering characteristics and semantic layout, without obvious artifacts and structural fractures. For large-format SAR images, the sliding window strategy realizes seamless stitching, ensuring the global spatial coherence and consistency of scattering characteristics of the whole image.

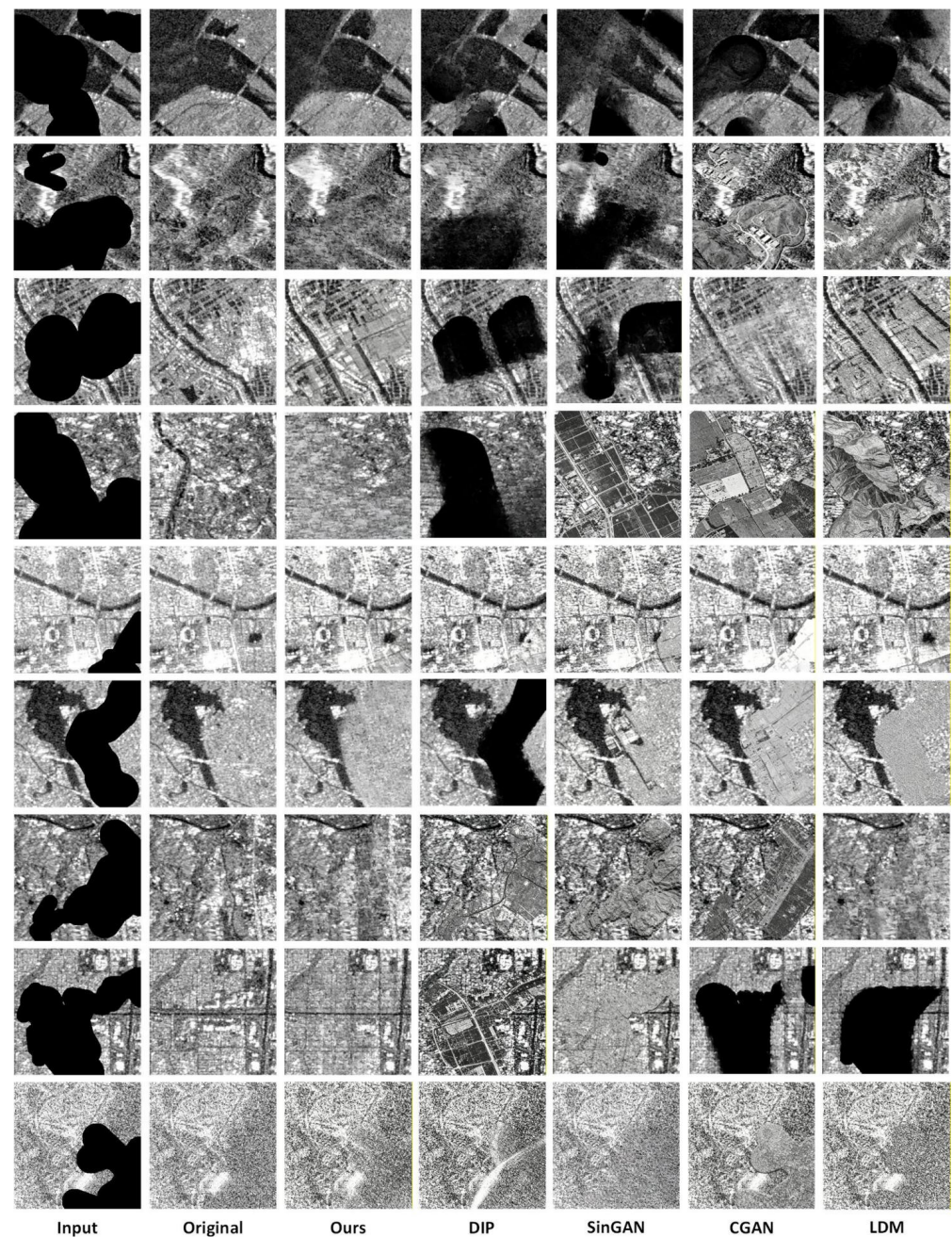


Figure 7. Qualitative comparison of SAR image inpainting results across different methods. From left to right: masked input, original image, the proposed method, DIP, SinGAN, CGAN, and LDM. Our method recovers detailed scattering patterns and maintains natural scene coherence, outperforming other methods that often produce over-smoothing, artifacts, or semantic inconsistencies in large missing regions.

We train the SAR-VAE using the full SAR dataset together with SAR images with random feature occlusion. After model convergence, the network is capable of encoding and reconstructing high-dimensional latent features of SAR imagery, as visualized in Figure 8. Figure 8 intuitively illustrates that the statistical distribution of images reconstructed by the model gradually approaches that of real SAR images with iterative training.

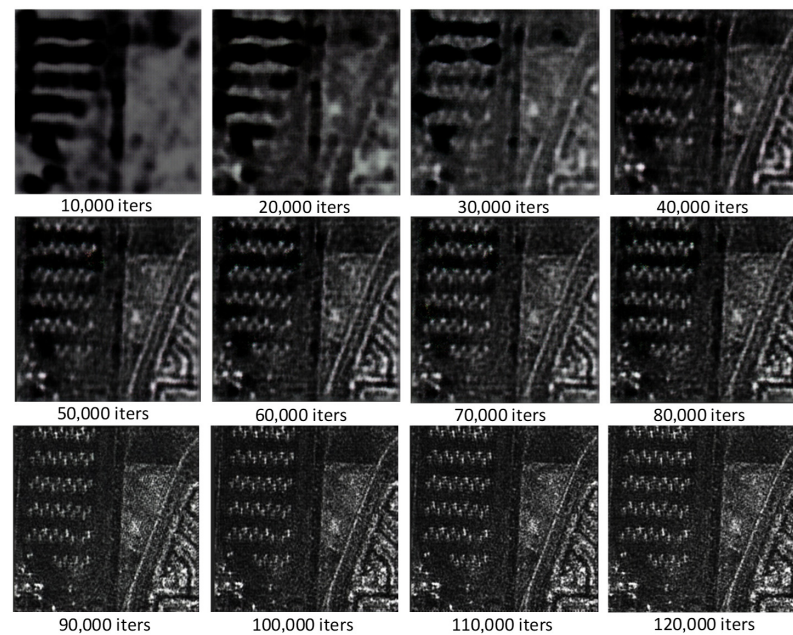


Figure 8. Schematic diagram of the generation effects during the training of SAR-VAE from 0 to 120,000 iterations. At the beginning, only blurred contours can be generated. As the number of training iterations increases, SAR-VAE can gradually generate clear textures.

3.4.5. Downstream Task Validation

To verify that the samples restored by the proposed method can effectively support the training of downstream tasks, this paper takes SAR image scene classification as the downstream task and carries out experiments using the MRSSC dataset.

In the experiment, the single-modal SAR data of the MRSSC dataset is selected, and the training set, validation set, and test set are evenly divided according to the ratio of 7:2:1. To simulate the data degradation problem in real SAR imaging, the training set images are uniformly processed with information missing, which is consistent with the mask setting in the previous sections. Then, each comparison method and the proposed FEDM are used to complete the inpainting and restoration, respectively.

ResNet50, the mainstream backbone network for remote sensing classification, is adopted. It is first trained on the original complete SAR images, and the Top-1 Overall Accuracy (OA) of 94.8% is obtained as the performance benchmark.

After that, ResNet50 is trained respectively on the training set images restored by different methods, and tested on the unified test set. With Top-1 OA as the core metric, the supporting capability of SAR images restored by each method for the classification task is quantitatively evaluated. The results are shown in Table 5.

Table 5. Performance of downstream SAR image scene classification.

Method	Top-1 OA
Original Images	0.948
Telea [2]	0.638
DIP [9]	0.708
CGAN [4]	0.797
SinGAN [5]	0.813
LDM [8]	0.831
Ours	0.917

The quantitative results in Table 5 show that the classification model trained on images restored by FEDM achieves a Top-1 overall classification accuracy of 0.917, reaching 96.7% of the classification performance obtained on original intact images, with a remarkable 8.6 percentage points improvement over the second-best basic LDM method.

This result further verifies that images restored by the proposed method can be applied to the training and inference of deep learning models for downstream tasks, and provides an effective solution to the problems of data missing and sample scarcity in practical SAR application scenarios.

4. Discussion

This study presents the Feature-Encoding Diffusion Model (FEDM) for large-scale SAR image inpainting, which addresses the core challenges of insufficient scattering fidelity and poor global semantic coherence by integrating a SAR-specific VAE, cross-attention semantic embedding, dual-condition guided diffusion, and an overlapping sliding-window strategy. Experiments on multiple datasets show that FEDM consistently outperforms state-of-the-art methods in PSNR, SSIM, ENL, and EPI, and downstream scene classification further confirms the practical reliability of the restored results.

The superior performance of FEDM mainly stems from its SAR-oriented design. The SAR-VAE, optimized with a multi-component joint loss, effectively captures the statistical characteristics of speckle and backscattering, ensuring physically plausible scattering reconstruction. The cross-attention semantic embedding injects scene-level priors into the diffusion process, alleviating semantic ambiguity and spatial layout disorder in single-channel SAR inpainting. The 50% overlapping sliding window balances quality and efficiency, enabling seamless, artifact-free restoration for large-format images. The ablation results in Section 3.4 further verify that each proposed module brings stable performance gains. The coordinated combination of SAR-VAE, cross-attention embedding and overlapping sliding window is the key to the overall superiority of our FEDM.

Compared with traditional interpolation, GAN-based methods, and vanilla diffusion models, FEDM shows clear advantages in both pixel accuracy and SAR-specific physical consistency. Quantitative results confirm that FEDM substantially outperforms LDM, achieving higher PSNR, SSIM, and ENL values. Downstream validation also demonstrates that the restored images preserve high-level semantic information, supporting reliable subsequent applications.

Nevertheless, this work has limitations. The model is validated on simulated missing masks and medium-resolution data; generalization to real noise/interference scenarios and ultra-high-resolution SAR images requires further verification. In addition, the sliding-window inference introduces moderate computational overhead. Future work will focus on lightweight diffusion architectures, multi-temporal consistency modeling, and unsupervised adaptation to improve practicality and generalization.

Overall, FEDM provides an effective solution for large-scale SAR image inpainting, which can enhance the quality and reliability of downstream remote sensing tasks such as target detection.

5. Conclusions

This paper proposes an effective method for large-scale SAR image inpainting, which is achieved by integrating two core modules into the state-of-the-art Latent Diffusion Model (LDM) framework. One is SAR-specific Variational Auto-Encoder (SAR-VAE), which focuses on accurately fitting the microwave backscattering distribution of ground features to improve scattering characteristic reconstruction quality. The other one is the cross-attention semantic embedding branch, which aims to inject scene semantic priors into

the diffusion process to ensure spatial layout rationality and global coherence of inpainting results. Experimental results on OS-dataset, SEN1-2 and SRSD-1.0 datasets demonstrate that our method significantly improves the baseline method LDM by large margins and achieves the state-of-the-art results compared with existing SAR image inpainting methods.

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

Funding: This research was funded by the National Science Foundation of China under Grant No. 62171040. The APC was funded by the same grant.

Data Availability Statement: The data presented in this study are derived from four publicly available remote sensing datasets: OSdataset, SEN1-2, SRSD-1.0 and MRSSC. All datasets are publicly accessible and properly cited in the manuscript.

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

References

- Kingma, D.P.; Welling, M. Auto-Encoding Variational Bayes. *arXiv* **2014**, arXiv:1312.6114. [\[CrossRef\]](#)
- Telea, A. An Image Inpainting Technique Based on the Fast Marching Method. *J. Graph. Tools* **2004**, *9*, 23–34. [\[CrossRef\]](#)
- Bhamidipati, S.R.M.; Srivatsa, C.; Gowda, C.K.S.; Vadada, S. Generation of SAR Images Using Deep Learning. *SN Comput. Sci.* **2020**, *1*, 355. [\[CrossRef\]](#)
- Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Munich, Germany, 5–9 October 2015; pp. 234–241.
- Mirza, M.; Osindero, S. Conditional Generative Adversarial Nets. *arXiv* **2014**, arXiv:1411.1784.
- Huang, H.; Zhang, F.; Zhou, Y.; Yin, Q.; Hu, W. High Resolution SAR Image Synthesis with Hierarchical Generative Adversarial Networks. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Yokohama, Japan, 28 July–2 August 2019; pp. 1–4. [\[CrossRef\]](#)
- Guo, J.; Lei, B.; Ding, C.; Zhang, Y. Synthetic Aperture Radar Image Synthesis by Using Generative Adversarial Nets. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 1111–1115. [\[CrossRef\]](#)
- Shaham, T.R.; Dekel, T.; Michaeli, T. SinGAN: Learning a Generative Model from a Single Natural Image. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4569–4579.
- Ulyanov, D.; Vedaldi, A.; Lempitsky, V. Deep Image Prior. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 9446–9454.
- Qosja, D.; Wagner, S.; O'Hagan, D. SAR Image Synthesis with Diffusion Models. In Proceedings of the IEEE Radar Conference, Denver, CO, USA, 6–10 May 2024. [\[CrossRef\]](#)
- Ying, Z.; Ke, W.; Zhai, Y.; Long, Z.; Zhou, J.; Zhu, H.; Chen, C.L.P. DiffuSAR: Frequency Domain-Aware Diffusion Model for SAR Image Generation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 11851–11866. [\[CrossRef\]](#)
- Moon, J.; Yun, J.; Kim, J.; Lee, J.; Kim, M. Diffusion-Based Data Augmentation and Knowledge Distillation with Generated Soft Labels Solving Data Scarcity Problems of SAR Oil Spill Segmentation. *arXiv* **2024**, arXiv:2412.08116.
- Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-Resolution Image Synthesis with Latent Diffusion Models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 10674–10685.
- Debussière, S.; Trounev, N.; Letheule, N.; Lévêque, O.; Colin, E. Quantitative Comparison of Fine-Tuning Techniques for Pretrained Latent Diffusion Models in the Generation of Unseen SAR Image Concepts. *arXiv* **2025**, arXiv:2506.13307.
- Xiang, Y.; Chen, J.; Hong, Z.; Jiao, N.; Wang, F.; You, H.; Tong, X. OSDataset2.0: SAR-Optical Image Matching Dataset and Evaluation Benchmark. *J. Radars* **2025**, *14*, 983–995.
- Schmitt, M.; Zhu, X.X. The SEN1-2 Dataset for Deep Learning in SAR-Optical Data Fusion. In *Proceedings of the ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, Karlsruhe, Germany, 10–12 October 2018*; Copernicus Publications: Göttingen, Germany, 2018; Volume IV-1, pp. 141–148.

17. Lei, S.; Lu, D.; Qiu, X.; Ding, C. SRSDD-v1.0: A High-Resolution SAR Rotation Ship Detection Dataset. *Remote Sens.* **2021**, *13*, 5104. [[CrossRef](#)]
18. Liu, K.; Yang, J.; Li, S. Remote-Sensing Cross-Domain Scene Classification: A Dataset and Benchmark. *Remote Sens.* **2022**, *14*, 4635. [[CrossRef](#)]
19. ITU-T. *Method for Objective Video Quality Assessment Using Peak Signal-to-Noise Ratio with Spatial and Temporal Alignment*; ITU-T Recommendation J.340; International Telecommunication Union (ITU): Geneva, Switzerland, 2009.
20. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [[CrossRef](#)] [[PubMed](#)]
21. Lee, J.S. Refined Filtering of Image Noise Using Local Statistics. *Comput. Graph. Image Process.* **1981**, *15*, 380–390. [[CrossRef](#)]
22. Sheng, Y.; Xia, Z.G. A Comprehensive Evaluation of Filters for Radar Speckle Suppression. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Lincoln, NE, USA, 31 May 1996; Volume 3, pp. 1559–1561.

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