MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

Multimodal Diffusion Bridge with Attention-Based SAR Fusion for Satellite Image Cloud Removal

Hu, Yuyang; Lohit, Suhas; Kamilov, Ulugbek; Marks, Tim K.

TR2025-138 October 01, 2025

Abstract

Deep learning has achieved some success in addressing the challenge of cloud removal in optical satellite images, by fusing with synthetic aperture radar (SAR) images. Recently, diffusion models have emerged as powerful tools for cloud removal, delivering higher-quality estimation by sampling from cloud-free distributions, compared to earlier methods. However, diffusion models suffer from limitations that can result in suboptimal performance. In particular, diffusion models initiate sampling from pure Gaussian noise, which complicates the sampling trajectory. Moreover, current methods often inadequately fuse SAR and optical data; simple concatenation of these disparate modalities at the input stage typically yields suboptimal results. To address these limitations, we propose Diffusion Bridges for Cloud Removal, DB-CR, which directly bridges between the cloudy and cloud-free image distributions. In addition, we propose a novel multimodal diffusion bridge architecture with a two-branch backbone for multimodal image restoration, incorporating an efficient backbone and dedicated cross-modality fusion blocks to effectively extract and fuse features from synthetic aperture radar (SAR) and optical images. By formulating cloud removal as a diffusion-bridge problem and leveraging this tailored architecture, DB-CR achieves high-fidelity results while being computationally efficient. We evaluated DBCR on the SEN12MS-CR cloud-removal dataset, demonstrating that it achieves state-of-the-art results.

IEEE Transactions on Geoscience and Remote Sensing 2025

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

Multimodal Diffusion Bridge with Attention-Based SAR Fusion for Satellite Image Cloud Removal

Yuyang Hu, Suhas Lohit, Ulugbek S. Kamilov, Tim K. Marks

***Abstract*—Deep learning has achieved some success in addressing the challenge of cloud removal in optical satellite images, by fusing with synthetic aperture radar (SAR) images. Recently, diffusion models have emerged as powerful tools for cloud removal, delivering higher-quality estimation by sampling from cloud-free distributions, compared to earlier methods. However, diffusion models suffer from limitations that can result in suboptimal performance. In particular, diffusion models initiate sampling from pure Gaussian noise, which complicates the sampling trajectory. Moreover, current methods often inadequately fuse SAR and optical data; simple concatenation of these disparate modalities at the input stage typically yields suboptimal results. To address these limitations, we propose Diffusion Bridges for Cloud Removal, DB-CR, which directly bridges between the cloudy and cloud-free image distributions. In addition, we propose a novel multimodal diffusion bridge architecture with a two-branch backbone for multimodal image restoration, incorporating an efficient backbone and dedicated cross-modality fusion blocks to effectively extract and fuse features from synthetic aperture radar (SAR) and optical images. By formulating cloud removal as a diffusion-bridge problem and leveraging this tailored architecture, DB-CR achieves high-fidelity results while being computationally efficient. We evaluated DB-CR on the SEN12MS-CR cloud-removal dataset, demonstrating that it achieves state-of-the-art results.**

***Index Terms*—Cloud Removal, Diffusion Bridge, Synthetic Aperture Radar, Multimodal Fusion**

I. Introduction

SATELLITE-BASED remote sensing imagery is important in a wide range of fields, such as agriculture monitoring [1] and land use mapping [2]. One of the biggest challenges is cloud coverage [3], which can decrease the image contrast and obscure important details.

To address this issue, there is growing interest in developing restoration algorithms to recover the lost or distorted signals of satellite images. Early methods for cloud removal in satellite imagery are based on the assumption that cloud-obscured areas share similar statistical and geometric characteristics with the cloud-free regions. By treating cloud removal as an image inpainting problem, these techniques aim to reconstruct missing data by utilizing information from neighboring uncorrupted areas [4].

Deep learning has emerged as a powerful tool for addressing cloud-removal tasks. Instead of explicitly defining an inpainting problem, deep learning methods employ deep neural networks (DNNs) to map cloud-covered images directly to their desired cloud-free[1] counterparts. However, due to the severely ill-posed nature of cloud removal (CR) problems, direct learning of such a mapping can lead to over-smooth predictions. To address this, generative models, particularly conditional Generative Adversarial Networks (cGANs), have shown the potential to produce realistic images. These include methods using SAR guidance with adversarial loss [5], [6], spatial-spectral feature alignment [7], and texture-preserving discriminators for cloud-free reconstruction [8], [9]. The training of cGANs relies on adversarial learning, in which a generator outputs estimated cloud-free images, and a discriminator tries to distinguish these estimates from ground-truth reference images. This adversarial training process helps cGANs to remove cloud coverage and produce images with reasonably realistic texture and details. However, one significant limitation of cGAN-based methods is the instability of their adversarial training process, which can lead to inconsistent performance.

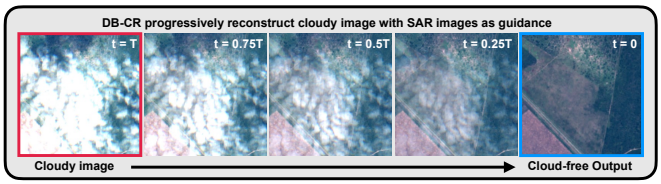

Fig. 1. Illustration of DB-CR's iterative process: leveraging SAR data and training a diffusion bridge approach to progressively refine cloud-covered images, reducing cloud cover step by step to produce a clear, cloud-free result.

Some studies explore multi-temporal CR methods, which use a sequences of images captured from the same location at different times. By leveraging the information of other cloud-free or cloudless time points to aid the reconstruction, multi-temporal CR methods achieve improved performance. In many applications, however, multi-temporal data may not be available or easily accessible.

Many cloud-removal techniques rely on synthetic aperture radar (SAR) data to enhance the reconstruction of cloud-obscured pixels, especially when only mono-temporal data (a single image from each location) are available. Prior studies include cGAN-based fusion [5]–[10], multi-scale or hierarchical fusion networks [11], [12], and uncertainty-aware reconstructions [13]. Since SAR is robust to weather conditions

Yuyang Hu is with the Department of Electrical & System Engineering, Washington University in St. Louis, USA. This work was done when Yuyang Hu was an intern at MERL.

Ulugbek S. Kamilov is with the Departments of Computer Science & Engineering and Electrical & System Engineering, Washington University in St. Louis, USA.

Suhas Lohit and Tim K. Marks are with Mitsubishi Electric Research Laboratories (MERL), USA.

[1]Throughout this paper, we use *cloud-free image*, *clean image*, and *target image* interchangeably to mean the ground-truth scene without cloud obstruction, noting that each term emphasizes a slightly different aspect depending on context.

and can penetrate clouds [14], SAR provides an important supplementary data stream to optical imagery for accurate cloud removal and surface reconstruction. In this work, we focus on multimodal data—SAR + optical images—which allows us to explore cloud removal in scenarios in which temporal information is limited.

Recently, diffusion models (DMs) have emerged as a promising alternative to cGAN-based methods for cloud removal [15]–[18]. DMs contain two Markov processes: a predefined forward process that progressively adds noise to clean data, and a learning-based reverse process that attempts to recover the original data from the noisy version. By iteratively adding and removing noise, DMs can learn and sample from the target distribution in a more stable manner. CR methods that use diffusion models represent the current state of the art in cloud removal [17], [18]. However, once a DM is trained, the inference is then limited to sample from the posterior distribution. According to the perception-distortion trade-off [19], this posterior sampling approach tends to optimize for perceptual quality at the expense of reconstruction accuracy (i.e., higher distortion). While this characteristic is advantageous for applications like text-to-image generation, where visual appeal is often prioritized, it is less suitable for cloud removal (CR). CR tasks are highly sensitive and demand outputs that faithfully represent the underlying scene, thus requiring methods that can enhance fidelity.

To address this fidelity challenge, Diffusion Bridges (DBs) present a promising alternative for image restoration tasks, such as image super-resolution or deblurring, where both initial (degraded) and target (clean) conditions are clearly specified. Existing DB methods include iterative direct inversion [20] and implicit-to-explicit score bridging [21]. Distinct from standard DMs, which learn to reverse a progressive noise-addition process, DBs directly model the transformation pathway between two specific, fixed states over time (see Figure 1). This direct modeling between defined endpoints addresses a significant limitation of standard DMs: their potential for insufficient control over the transition path, which can lead to oversmoothed or inconsistent results. Consequently, by constraining the model to operate between these specific states, DBs demonstrably enhance reconstruction stability and fidelity. We believe that these intrinsic characteristics—direct state-to-state modeling leading to enhanced fidelity and stability—are well-suited to the cloud removal problem, where the initial (cloudy) and target (cloud-free) states are clearly defined.

Despite this strong theoretical suitability for CR, their application to this specific task has not previously been explored. Consequently, their specific efficacy and advantages for CR tasks have not been evaluated. Furthermore, existing DB methods do not incorporate multimodal inputs. This is a notable omission, as auxiliary data, such as Synthetic Aperture Radar (SAR) imagery, can provide crucial prior information essential for robust CR performance, especially in scenes with heavy cloud cover.

To bridge these research gaps, we propose our multimodal Diffusion Bridge for Cloud Removal (DB-CR). Our work makes four key contributions:

1) We introduce the *first* application of diffusion bridges to the cloud removal task, DB-CR. Our method, which is specifically designed for multimodal cloud removal, is more stable and accurate than existing approaches in literature, outperforming even the latest methods that are based on diffusion models.
2) We propose a novel multimodal diffusion bridge architecture, incorporating a two-branch time-embedded feature extraction block and a cross-modal attention-based fusion block to effectively leverage information from both modalities—optical and SAR images—for effective cloud removal.
3) We perform extensive experiments, which clearly show that DB-CR achieves state-of-the-art performance compared with several existing cloud-removal methods on the widely used SEN12MS-CR benchmark.
4) Our method offers the flexibility to adjust the number of inference steps, enabling a customizable trade-off between minimizing a distortion metric like mean-squared error that leads to blurrier outputs and maximizing perceptual performance by preserving sharper details like roads and rivers in the recovered cloud-free optical image[2].

II. Background

A. Conditional diffusion models for image restoration

Conditional Diffusion Models (cDMs) are designed to learn a parameterized Markov chain to convert a Gaussian distribution into a conditional data distribution $p(\boldsymbol{x}|\boldsymbol{c})$ [22]. Recently, many cDM-based methods have shown great performance in various image restoration tasks [23]. In the multimodal cloud-removal task, $\boldsymbol{x}$ refers to the target cloud-free optical image, and the condition $\boldsymbol{c}$ refers to the combination of the corresponding cloudy image and SAR image. The forward diffusion process starts from a clean sample $\boldsymbol{x}_0 := \boldsymbol{x} \sim p(\boldsymbol{x})$, then gradually adds Gaussian noise to $\boldsymbol{x}_0$ according to the following transition probability:

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) := \mathcal{N}(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t\mathbf{I}), \tag{1}$$

where $t \in [0, T]$, $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the Gaussian probability density function (pdf) over $\boldsymbol{x}$ with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and β_t is the noise schedule of the process. By the parameter changes $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$, we can rewrite $\boldsymbol{x}_t$ as a linear combination of $\boldsymbol{\epsilon}_t$ and $\boldsymbol{x}_0$:

$$\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t, \tag{2}$$

where $\boldsymbol{\epsilon_t} \sim \mathcal{N}(0, \mathbf{I})$. This yields a closed-form expression for the marginal distribution of $\boldsymbol{x}_t$ given $\boldsymbol{x}_0$, which simplifies sampling across multiple steps of the Markov chain:

$$q(\boldsymbol{x}_t|\boldsymbol{x}_0) := \mathcal{N}(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\mathbf{I}). \tag{3}$$

The goal of the reverse process is to generate a clean image $\boldsymbol{x}_0$ given a pure Gaussian noise image $\boldsymbol{x}_T$ and the condition $\boldsymbol{c}$. This is achieved by training a neural network $\boldsymbol{\epsilon_\theta}$, parameterized by $\boldsymbol{\theta}$, to reverse the Markov Chain from $\boldsymbol{x}_T$ to $\boldsymbol{x}_0$ one

[2]This trade-off is present in all image restoration algorithms [19], but using diffusion bridges allows customizing the trade-off point easily.

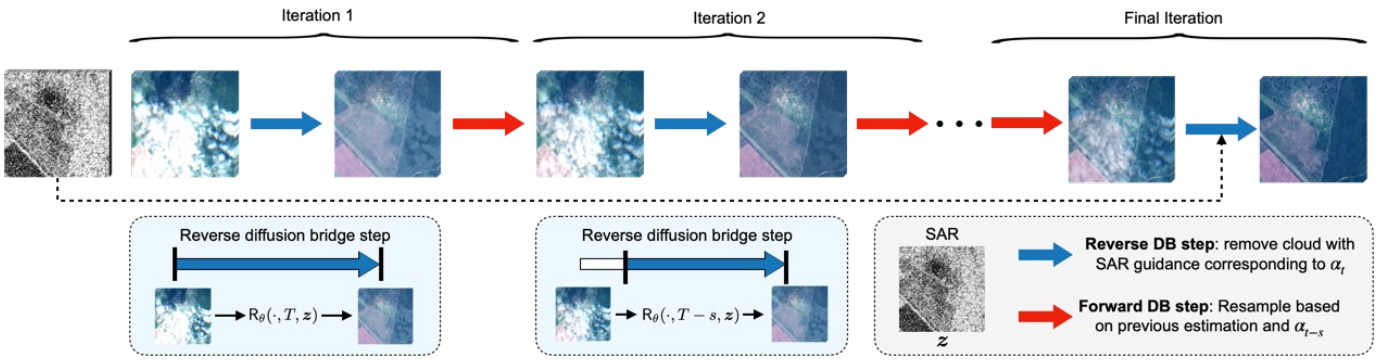

Fig. 2. A schematic diagram of our method, illustrating the inference process. Each iteration alternates between a reverse diffusion-bridge step, which estimates the cloud-free state, and a forward diffusion-bridge step, which progresses the state to the next diffusion timestep. The process starts from the cloudy image, progressively refining it toward a cloud-free output. T is the number of diffusion timesteps used in training the diffusion bridge. During inference, $N = \frac{T}{s}$ diffusion timesteps are used, where s is the step size. $\boldsymbol{z}$ refers to the SAR image.

step at a time. The learning is formulated as the estimation of a parameterized Gaussian process

$$p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) := \mathcal{N}(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t, \boldsymbol{c}), \sigma_t^2 \mathbf{I}), \tag{4}$$

where the mean $\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t, \boldsymbol{c})$ is only variable we need to estimate, and $\sigma_t^2 = \frac{1-\bar{\alpha}_t}{1-\bar{\alpha}_{t-1}}\beta_t$. In cDM, $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}$ is trained to predict the noise $\boldsymbol{\epsilon}_{\boldsymbol{t}}$ added during the forward process by minimizing the following expected loss:

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\epsilon}_{\boldsymbol{t}}, \boldsymbol{x}_t, t}\left[\|\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t, \boldsymbol{c}) - \boldsymbol{\epsilon}_{\boldsymbol{t}}\|_2^2\right]. \tag{5}$$

The estimated noise is then used to reparameterize the sampling from the conditional distribution:

$$\boldsymbol{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t, \boldsymbol{c})\right) + \sigma_t \boldsymbol{z}, \tag{6}$$

where $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. By iteratively applying this reverse process from $t = T$ to $t = 0$, the trained cDM can reconstruct the clean sample $\boldsymbol{x}_0$ by progressively sampling from the learned conditional distribution.

B. Conditional diffusion models for cloud removal

Recent works have employed cDMs to address cloud removal tasks, leveraging the power of diffusion models to produce high-quality cloud-free images, such as in DDPM-CR [24] and DiffCR [17]. In DDPM-CR, the authors directly applied cDMs [22] to the cloud removal problem by concatenating the cloudy image as a condition with the input of the diffusion model. DiffCR further advanced this by introducing a more efficient network architecture, incorporating an additional encoder to better utilize conditional inputs, and achieving faster inference through an optimized ODE solver rather than the original DDPM solver. These cDM-based methods have demonstrated substantial improvements in reconstructing cloud-free satellite imagery, particularly in scenarios with varying degrees of cloud coverage.

However, despite promising results, the stochastic nature of the reverse process can sometimes lead to inconsistencies and artifacts, especially in areas where the model needs to be more precise in reconstructing details. This motivated us to consider the use of more deterministic approaches. In particular, we focused on the diffusion bridge framework. In our approach, the diffusion bridge explicitly controls the transformation between the cloudy and cloud-free states, providing a more stable and effective solution to the cloud removal problem.

C. Diffusion bridges for image restoration

Generation in DMs is achieved by transforming noise into a target distribution through a series of denoising steps. The diffusion bridge (DB) method generalizes this approach by modeling a stochastic process that connects two fixed states, where the initial state is not limited to a Gaussian distribution. In the context of image restoration, DBs often involve constructing a probabilistic trajectory between two states (e.g., degraded and clean images) using an optimal transport formulation [20], [21], [25]. This approach leverages the inherent similarity between these states to efficiently learn mappings, leading to improved performance and faster inference. For example, InDI [20] proposed a diffusion bridge framework that achieved state-of-the-art results across various real-world image restoration tasks, including image super-resolution, image deblurring, and JPEG artifact removal, while achieving significantly faster performance compared to traditional DMs. Moreover, recent work [26] demonstrates that a DB-based restoration model offers superior robustness to distributional shifts between training and testing datasets compared with DM-based method. By learning to restore a diverse range of degraded states, DB models provide enhanced stability to out-of-distribution data. While DBs have demonstrated strong performance in single modality settings, it has neither been extended to multimodal restoration tasks nor applied to the specific challenges of cloud removal.

A diffusion bridge explicitly models the transition from a degraded image state to a clean image state through an Optimal Transport (OT) framework. Given a tuple $(\boldsymbol{x}_0, \boldsymbol{y}) \sim p(\boldsymbol{x}, \boldsymbol{y})$, where $\boldsymbol{y}$ represents the degraded image and $\boldsymbol{x}_0$ represents the target clean image, the forward process in DB begins from $\boldsymbol{x}_0$ and transforms it towards $\boldsymbol{y}$, with each intermediate state $\boldsymbol{x}_t$ representing a linear mixture of both the corrupted and target

images. This process is defined as:

$$\boldsymbol{x}_t = (1-\alpha_t)\boldsymbol{x}_0 + \alpha_t \boldsymbol{y}, \quad \alpha_t \in [0,1], \tag{7}$$

where α_t is a monotonically increasing function as t goes from 0 to T. Specifically, $\alpha_0 = 0$ and $\alpha_T = 1$. As t progresses from 0 to T in the forward process, $\boldsymbol{x}_t$ transitions smoothly from the clean image $\boldsymbol{x}_0$ to the degraded image $\boldsymbol{y}$, creating a well-defined trajectory that deterministically blends the two image states. Conversely, in the reverse process, as t progresses from T to 0, $\boldsymbol{x}_t$ evolves from $\boldsymbol{y}$ back to $\boldsymbol{x}_0$ through a deterministic trajectory, progressively removing the degradation and reconstructing the original clean image.

The reverse process in the DB framework reconstructs the clean image from the degraded observation through a multi-step sampling strategy, utilizing an Optimal Transport Ordinary Differential Equation (OT-ODE). Unlike traditional end-to-end cloud removal models that attempt to estimate a clean image in a single pass, the Diffusion Bridge breaks down the transformation into a sequence of smaller, more manageable steps, each guided by the OT-ODE. The OT-ODE is described by:

$$\frac{d\boldsymbol{x}_t}{dt} = v_t(\boldsymbol{x}_t|\boldsymbol{x}_0), \quad \text{where} \quad v_t(\boldsymbol{x}_t|\boldsymbol{x}_0) = \alpha_t(\boldsymbol{x}_0 - \boldsymbol{x}_t), \tag{8}$$

where $v_t(\boldsymbol{x}_t|\boldsymbol{x}_0)$ represents the optimal velocity field that guides $\boldsymbol{x}_t$ towards the target state $\boldsymbol{x}_0$. To approximate this velocity field, a neural network $\mathsf{R}_\theta(\boldsymbol{x}_t, t)$ is trained to determine the optimal direction for each step of the reverse evolution. The training objective for the network is formulated as:

$$\theta^* = \arg\min_\theta \mathbb{E}_{t,\boldsymbol{x}_t,\boldsymbol{x}_0}\left[\|\mathsf{D}_\theta(\boldsymbol{x}_t, t) - v_t(\boldsymbol{x}_t|\boldsymbol{x}_0)\|_2^2\right], \tag{9}$$

where $\mathsf{D}_\theta(\boldsymbol{x}_t, t)$ predicts the optimal velocity that guides $\boldsymbol{x}_t$ back to the clean target state $\boldsymbol{x}_0$. The multi-step reverse sampling provides incremental updates toward the optimal transport solution, minimizing transport costs and preserving structural details. The controlled velocity at each stage ensures a smooth transition, reducing artifacts and oversmoothing, which are commonly observed in ill-posed problems. By decomposing the reconstruction into multiple steps, the diffusion bridge framework addresses the complexity of difficult restoration tasks by solving progressively simpler subproblems, showing state-of-the-art performance in various image restoration tasks (e.g., image super-resolution, image deblurring, and image inpainting).

Existing diffusion bridge frameworks have primarily been developed for single-modality image restoration tasks. They have not been applied to or validated on challenging multimodal problems such as cloud removal. In this work, we introduce a new multimodal diffusion brige framework, DB-CR, which is specifically tailored for cloud removal. Not only is the first time that a diffusion bridge has been applied to the task of cloud removal, but as far as we know it is also the first multimodal diffusion bridge. DB-CR incorporates a novel attention fusion mechanism and an efficient dual-branch backbone to effectively integrate structural information from SAR data with fine spectral details from optical imagery.

Unlike prior methods, which lack the capacity to handle multimodal inputs, DB-CR leverages the complementary strengths of SAR for structural information and optical imagery for fine details, achieving state-of-the-art performance. By demonstrating DB-CR on cloud removal—a highly ill-posed multimodal problem—this work sets a new benchmark and establishes the utility of diffusion bridges for previously unexplored multimodal domains.

In the following sections, we detail the innovations in DB-CR and demonstrate that its multimodal diffusion bridge formulation advances the state of the art cloud-removal performance.

III. Proposed Method

A. Problem definition

This work focuses on the SAR-conditioned cloud removal problem. The objective is to generate a cloud-free optical satellite image, $\boldsymbol{x} \in \mathbb{R}^{C,H,W}$, from a single observed cloudy optical image, $\boldsymbol{y} \in \mathbb{R}^{C,H,W}$, and the corresponding SAR image $\boldsymbol{z} \in \mathbb{R}^{C',H,W}$. In our experiments, $C = 13$ is the number of multispectral channels in each optical image, $C' = 2$ is the number of channels in the SAR image, and H and W are the images' height and width, respectively.

B. Multimodal Diffusion ridge framework for Cloud Removal (DB-CR)

Given a pair of spatially aligned optical images, one cloudy and one cloud-free, we define the forward process of the diffusion bridge (DB) at each discrete time step as

$$\boldsymbol{x}_t = (1-\alpha_t)\boldsymbol{x}_0 + \alpha_t \boldsymbol{y}, \quad t \in \{0, 1, \ldots, T\}, \tag{10}$$

where $\boldsymbol{x}_t$ represents an intermediate image transitioning from the cloud-free image $\boldsymbol{x}$ (at $t = 0$) to the cloudy image $\boldsymbol{y}$ (at $t = T$). The weighting factor α_t is a function that increases monotonically from 0 to 1 as t goes from 0 to T, ensuring a smooth and deterministic trajectory between the two image states.

The reverse process aims to recover the clean image $\boldsymbol{x}$ from the cloudy image $\boldsymbol{y}$ by inverting this DB forward process. To accomplish this, we train a neural network $\mathcal{R}_\theta$ to predict the clean image at each time step t, given the corrupted intermediate image $\boldsymbol{x}_t$ and the aligned SAR reference image $\boldsymbol{z}$. The training objective is defined as:

$$\theta^* = \arg\min_\theta \mathbb{E}_{t\sim\{0,1,\ldots,T\}} \left\|\mathsf{R}_\theta(\boldsymbol{x}_t, t, \boldsymbol{z}) - \boldsymbol{x}_0\right\|_1, \tag{11}$$

where the model directly learns to predict $\boldsymbol{x}_0$ at each time step. Here, we use the mean absolute error (MAE) loss function, which has demonstrated good performance for the cloud removal task [11]. In this formulation, R_θ learns to predict the minimum mean absolute error (MMAE) estimation, $\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_t]$, of true cloud-free image $\boldsymbol{x}_0$ from the intermediate corrupted input $\boldsymbol{x}_t$. The pseudocode for this training process is provided in Algorithm 1.

Once the network is trained, we define a reverse process to reconstruct the cloud-free image $\boldsymbol{x}_0$ from the cloudy observation $\boldsymbol{x}_T$. As t is decreased from T (initial step) to 0 (final step), α_t progressively decreases from 1 to 0, and the reverse

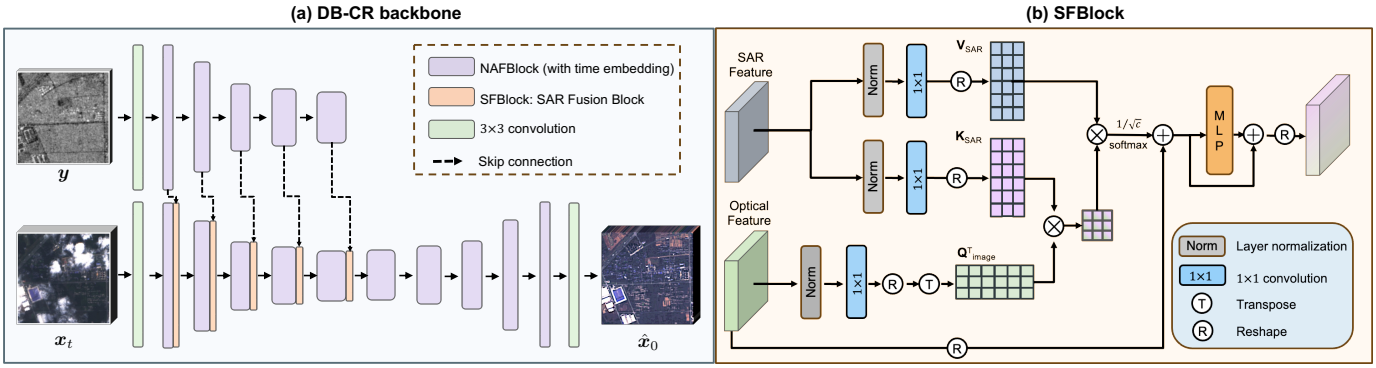

Fig. 3. The architecture of the DB-CR backbone network. (a) This two-branch design combines a U-Net restoration branch (bottom), which utilizes NAFBlocks (purple rectangles) for efficient restoration, and a SAR feature extraction branch (top). The network utilizes SFBlocks (orange rectangles) for feature fusion across branches. (b) Detailed architecture of each SFBlock.

process incrementally refines the image, transforming $\boldsymbol{x}_T$ back to $\boldsymbol{x}_0$.

Given a predefined number of function evaluations (NFE), we uniformly select N timesteps from the original diffusion process, by setting an appropriate step size s. Following this schedule of diffusion timesteps, the process starts with the original cloudy image $\boldsymbol{y}$ as the initial input and iteratively transforms it toward a cloud-free state. At each iteration, the pre-trained model R_θ computes the cloud-free estimate $\boldsymbol{x}'_{0|t}$ based on the current intermediate image $\boldsymbol{x}_t$, time step t, and auxiliary SAR data $\boldsymbol{z}$. Using Equation (12), the next intermediate image $\boldsymbol{x}_{t-1}$ is resampled by combining $\boldsymbol{x}'_{0|t}$ with $\boldsymbol{x}_t$.

$$\boldsymbol{x}_{t-s} = \left(1 - \frac{\alpha_{t-s}}{\alpha_t}\right)\boldsymbol{x}'_{0|t} + \left(\frac{\alpha_{t-s}}{\alpha_t}\right)\boldsymbol{x}_t, \quad s < t, \qquad (12)$$

where s is the step size of the reverse process and $\boldsymbol{x}'_{0|t}$ represents the cloud-free prediction based on the intermediate image $\boldsymbol{x}_t$, i.e., $\boldsymbol{x}'_{0|t} = \mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_t]$. The overall inference pseudocode is given in Algorithm 2. As illustrated in Figure 2, this repeated resampling progressively reduces cloud cover, ultimately producing the final cloud-free image.

The multi-step generation approach effectively addresses the challenge of over-smoothing, which is common in ill-posed problems like cloud removal. Unlike single-step methods that directly estimate cloud-free images and often produce averaged outputs lacking fine details, this approach decomposes the task into a series of simpler sub-problems under a diffusion bridge formulation. At each step, the model incrementally refines a less degraded image, progressively removing the cloud cover. This iterative refinement not only mitigates over-smoothing but also preserves finer details. Additionally, the stepwise process allows for more effective integration of SAR information at each stage, leading to sharper and more accurate cloud-free reconstructions.

C. SAR-enhanced diffusion bridge backbone

To enhance cloud removal performance and improve feature extraction from SAR data, we propose a novel two-branch diffusion bridge backbone, consisting of a SAR feature extraction branch and an optical image restoration branch, as illustrated in Figure 3. The backbone is built using two main components: (1) a time-embedded Nonlinear Activation-Free Network (NAFNet) block, and (2) a cross-attention based SAR Fusion Block (SFBlock). The SFBlock enables multi-level fusion of features from both the SAR extraction branch and the optical image restoration branch.

1) Two-branch backbone for DB-CR: As shown in Figure 3, the proposed DB-CR backbone has two branches: a NAFNet-based U-net architecture for image restoration (bottom row) and a NAFNet encoder architecture for SAR feature extraction (top row). This dual-branch design enables dedicated feature extraction for each modality, allowing the network to more efficiently learn modality-specific features.

Traditional diffusion bridge architectures often employ the U-Net using residual blocks and attention mechanisms such as channel-attention and self-attention. While effective in capturing fine details, these components significantly increase

Algorithm 1 DB-CR (Training)

Require: Training dataset: $p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$, Training diffusion timesteps T, $\{\alpha_t\}_{t=0}^{T}$

repeat:
  $\boldsymbol{x}_0, \boldsymbol{y}, \boldsymbol{z} \sim p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$, $t \sim \{0, 1, \ldots, T\}$
  $\boldsymbol{x}_t = (1 - \alpha_t)\boldsymbol{x}_0 + \alpha_t \boldsymbol{y}$
  Take gradient descent step on
    $\nabla_\theta \left\| \mathsf{R}_\theta\left(\boldsymbol{x}_t, t, \boldsymbol{z}\right) - \boldsymbol{x}_0 \right\|_1$
until convergence

Algorithm 2 DB-CR (Inference)

Require: Cloudy optical and SAR images: $(\boldsymbol{y}, \boldsymbol{z})$, Trained DB-CR model R_θ,Training diffusion timesteps T, NFE steps N

Get step size: $s = \frac{T}{N}$
Initial: $\hat{\boldsymbol{x}}_T = \boldsymbol{y}$
for $t = T$ to 0 with step size s **do**
  $\boldsymbol{x}'_{0|t} = \mathsf{R}_\theta(\boldsymbol{x}_t, t, \boldsymbol{z})$
  $\boldsymbol{x}_{t-s} \leftarrow \left(1 - \frac{\alpha_{t-s}}{\alpha_t}\right)\boldsymbol{x}'_{0|t} + \frac{\alpha_{t-s}}{\alpha_t}\boldsymbol{x}_t$
end for
return $\boldsymbol{x}'_{0|t}$

computational cost, which can slow down the inference process, particularly in iterative, multi-step generation required by diffusion bridge methods.

To address this drawback, we adopt a more efficient approach by incorporating Nonlinear Activation-Free Blocks (NAFBlocks) [25], [27] into our architecture, thus introducing a more efficient alternative to the commonly used ResBlock backbone. Typically, diffusion bridge frameworks for image restoration, as seen in [21], employ a ResBlock backbone that combines residual and attention blocks to enhance feature representation. However, this setup can be computationally expensive due to the added complexity of attention mechanisms and activation layers. In contrast, NAFBlocks take a streamlined approach by excluding attention mechanisms and activation functions, focusing solely on linear transformations, which significantly reduces the computational overhead. Each NAFBlock consists of two parts: (1) a mobile convolution module (MBConv) based on point-wise and depth-wise convolution with channel attention, and (2) a feed-forward network (FFN) module that has two fully connected layers. The LayerNorm (LN) layer is added before both MBConv and FFN, and a residual connection is employed for both modules. We denote the intermediate feature maps as $\boldsymbol{Z}_{\text{MBConv}}$ and $\boldsymbol{Z}_{\text{FFN}}$. For a input feature map $\mathbf{X}$, the whole process is formulated as:

$$\begin{aligned} \boldsymbol{Z}_{\text{MBConv}} &= \text{MBConv}(\text{LN}(\mathbf{X})) + \mathbf{X} \\ \boldsymbol{Z}_{\text{FFN}} &= \text{FFN}(\text{LN}(\boldsymbol{Z}_{\text{MBConv}})) + \boldsymbol{Z}_{\text{MBConv}} \end{aligned} \tag{13}$$

Another special part of NAFBlocks is that they use SimpleGate units to replace nonlinear activation functions (such as ReLU or GELU) after the FFN module. Given an input feature map generated by FFN module $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$, SimpleGate first splits the input into two features $\mathbf{X}_1, \mathbf{X}_2 \in \mathbb{R}^{H\times W\times C/2}$ along the channel dimension. It then computes the output using element-wise multiplicative gating:

$$\text{SimpleGate}(\mathbf{X}) = \mathbf{X}_1 \odot \mathbf{X}_2 \tag{14}$$

While NAFBlocks were originally designed for general RGB image restoration, their ability to streamline feature extraction without sacrificing performance makes them particularly suitable for our multi-spectral cloud removal task. By introducing the NAFBlocks, we reduce the reliance on heavy attention mechanisms and activation functions, achieving a balance between efficiency and accuracy in cloud-free image prediction. This high-efficiency block addresses the fundamental challenge of high computational demand in multi-step restoration, which is crucial in diffusion bridges, resulting in faster inference while preserving image quality.

2) SAR-image cross-modal attention fusion: To effectively fuse the multimodal information from both the SAR and optical image branches, we employ a cross-modal attention mechanism in the SAR Fusion Block (SFBlock), inspired by the network structures in [28], [29]. This block enables the model to capture the complementary nature of SAR and optical data by aligning and integrating features at multiple levels. SAR data, which is robust to atmospheric conditions, provides structural and texture information, while the optical data contributes color and fine details. By employing multi-head cross-attention between these modalities, the SFBlock ensures that the final reconstruction benefits from the strengths of both data sources, ultimately leading to improved cloud removal and image restoration performance.

The SFBlock operates by applying a cross-modal attention mechanism similar to self-attention but with a key difference: the queries ($\mathbf{Q}_{\text{O}}$) are derived from the optical image branch, while the keys ($\mathbf{K}_{\text{S}}$) and values ($\mathbf{V}_{\text{S}}$) come from the SAR branch. In this setup, h and w represent the height and width of the feature maps, and c denotes the number of channels. Initially, both input features have dimensions $h \times w \times c$, but after passing through normalization and 1×1 convolution layers, they are reshaped to $hw \times c$, where hw represents the flattened spatial dimensions. The cross-modal attention mechanism selectively emphasizes relevant SAR features using guidance from the optical image content, leveraging the unique characteristics of each modality. For each attention head, this process is expressed as:

$$\text{Attention}(\mathbf{Q}_{\text{O}}, \mathbf{K}_{\text{S}}, \mathbf{V}_{\text{S}}) = \mathbf{V}_{\text{S}} \cdot \text{softmax}\left(\frac{\mathbf{Q}_{\text{O}}^T \mathbf{K}_{\text{S}}}{\sqrt{c}}\right), \tag{15}$$

where $\mathbf{Q}_{\text{O}}^{\text{T}}\mathbf{K}_{\text{S}} \in \mathbb{R}^{c\times c}$ represents the similarity scores between optical and SAR features. These scores are computed by taking the transpose of the queries, $\mathbf{Q}_{\text{O}}^{\text{T}}$ (dimension $c \times hw$), and multiplying it by $\mathbf{K}_{\text{S}}$ (dimension $hw \times c$), yielding a matrix in $\mathbb{R}^{c\times c}$. The resulting attention map in $\mathbb{R}^{c\times c}$ is then applied to $\mathbf{V}_{\text{S}}$ (dimension $hw \times c$) to produce the final attended features in $\mathbb{R}^{hw\times c}$. This channel-wise attention significantly reduces spatial complexity from $O(h^2w^2)$ to $O(c^2)$, making the operation more efficient without sacrificing performance. Finally, the output of the attention operation is added to the input image features to produce $\boldsymbol{Z}_{\text{sum}}$, which is subsequently passed through a residual-connected multi-layer perceptron (MLP):

$$\boldsymbol{Z}_{\text{output}} = \boldsymbol{Z}_{\text{sum}} + \text{MLP}(\boldsymbol{Z}_{\text{sum}}). \tag{16}$$

After reshaping $\boldsymbol{Z}_{\text{output}}$ back to $c \times h \times w$, it serves as the final output of each attention head within the SFBlock. To produce the final output of the entire SFBlock, the outputs from all attention heads are concatenated along the channel dimension. A final 1×1 convolution layer is then applied as a linear projection to integrate information across the heads and ensure that the output tensor retains the same channel dimension as the input.

By integrating the efficiency of NAFBlocks with the advanced multimodal fusion capabilities of the SFBlock, our two-branch diffusion bridge backbone is designed for effective cloud removal. This dual-branch architecture combines the structural strengths of the SAR feature extraction branch with the fine details from the optical image restoration branch. Our approach addresses the computational challenges of multi-step restoration while ensuring accurate cloud-free reconstructions that preserve both spatial and spectral integrity.

IV. Experiments

A. Dataset

We conduct our experiments using the SEN12MS-CR dataset [30], a multimodal, mono-temporal dataset specifically

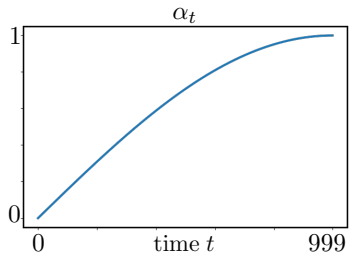

Fig. 4. α_t scheduling for DB-CR, which follows sine-based distribution.

curated for cloud removal tasks in remote sensing. SEN12MS-CR provides paired and co-registered data from SAR and multi-spectral optical imagery, collected from the Sentinel-1 and Sentinel-2 satellites of the European Space Agency's Copernicus mission respectively. The optical data contain 13 spectral bands, including visible (RGB), near-infrared (NIR), and shortwave infrared (SWIR) bands. The SAR data contain 2 channels containing VV and VH polarized images. The dataset contains observations covering 175 globally distributed regions of interest recorded in one of four seasons throughout the year of 2018. The full images are divided into a total of 122,218 patch triplets, each patch of size 256×256. The training, validation, and test sets contain 114056, 7176, and 7899 examples, respectively, following the dataset splits provided in [13].

Before the images are input into the network, the images undergo preprocessing steps to ensure quality and numerical stability, which are the same as in [31]. The Sentinel-2 optical bands are clipped to the range [0, 10,000], then scaled to the range [0, 1]. Similarly, the SAR data are clipped, with VV and VH polarizations being clipped to the range [-25, 0] and [-32.5, 0], respectively. To further align the SAR and optical data distributions, the Sentinel-1 SAR values are shifted into the positive domain and scaled to the range [0, 1], matching the scaled optical data values.

B. Implementation

All experiments were implemented using PyTorch 1.13.1 with CUDA 12.6. For training, we set the number of timesteps for the diffusion bridge process to 1000, and the scheduling of α_t corresponding to each timestep is the 'Sine Schedule' as shown in Figure 4. The training process was conducted over 50 epochs with a batch size of 4, using the Adam optimizer with a learning rate of 5×10^{-5}. These training parameters are chosen based on validation set performance. We evaluate both the distortion and perceptual performance of our proposed method, as well as several baseline methods. In the backbone architecture, the SAR and optical branches adopt channel dimensions of [22, 44, 88, 176] at each level. The encoder for both branches is configured with [1, 1, 1, 28] NAFBlocks across levels, while the decoder in the optical branch is designed with [1, 1, 1, 1] NAFBlocks. In both branches, before the NAFBlocks a single 1×1 convolutional layer is used which maps from 13 channels to 22 channels for the optical image branch, and from 2 channels to 22 channels for the SAR image branch. The outputs from each level of the encoder in both branches serve as inputs to the SFBlock. The SFBlock utilizes an increasing number of attention heads across levels, configured as [1, 1, 2, 4]. The MLP inside consists of two fully connected layers with GELU activation functions, where the hidden dimension is configured to be twice the size of the input dimension.

C. Evaluation metrics

For evaluating the restoration quality of the reconstructed cloud-free images, we employ a set of widely used quantitative metrics: Peak Signal-to-Noise Ratio (PSNR, in dB), Structural Similarity Index Measure (SSIM), Mean Absolute Error (MAE), and Spectral Angle Mapper (SAM). PSNR measures the pixel-wise similarity by comparing the reconstructed image to the ground truth, with higher values indicating less distortion. SSIM captures structural similarity by evaluating structural information and luminance consistency. MAE calculates the average absolute difference between predicted and true pixel values, highlighting overall error, while SAM, measured in degrees, quantifies angular deviation in spectral information, important for preserving spectral fidelity. These metrics are calculated using the code provided by UnCRtainTS[3] [13].

To measure perceptual performance, we use Learned Perceptual Image Patch Similarity (LPIPS) and Fréchet Inception Distance (FID). LPIPS is a perceptual metric that assesses similarity based on deep feature activations from neural networks, offering insight into the perceptual closeness of images. FID compares the distribution of deep features between sets of generated and real images, with lower values indicating closer alignment to the original image distribution. Both LPIPS and FID are essential for evaluating perceptual realism, as they reflect how closely reconstructed images resemble ground truth when viewed by a human observer. As the backbone neural networks for both LPIPS and FID are trained on RGB images, we compute these metrics exclusively on the RGB channels of the cloud-free estimations that are generated by the baseline methods and our proposed approach.

To measure model efficiency, we report the parameter count (Params), which represents the model's size and potential memory footprint, and the number of multiply–accumulate operations (MACs), a measure of computational complexity. These values provide insight into model scalability and resource requirements, particularly useful for deployment in resource-constrained environments. Calculations are performed using the open-source THOP repository[4].

D. Comparison with baseline methods

We selected several well-known methods as references to compare the performance of DB-CR:

- **SAR-Opt-cGAN** [8]: SAR-Opt-cGAN is a conditional GAN based method which uses SAR images as a prior to guide the reconstruction network.
- **DSen2-CR** [11]: Dsen2-CR method removes clouds from Sentinel-2 satellite images using a deep residual neural

[3]https://github.com/PatrickTUM/UnCRtainTS
[4]https://github.com/ultralytics/thop

TABLE I
COMPARISON OF DB-CR (USING NFE=1) WITH STATE-OF-THE-ART CLOUD-REMOVAL METHODS ON SEN12MS-CR DATASET. THE RESULTS CLEARLY SHOW THAT DB-CR ACHIEVES BEST PERFORMANCE WITH RESPECT TO ALL THE RECONSTRUCTION EVALUATION METRICS WHILE BEING COMPUTATIONALLY EFFICIENT.

Methods	PSNR(dB) ↑	SSIM ↑	MAE ↓	SAM (deg) ↓	FID ↓	LPIPS ↓	Parameters ($\times 10^6$) ↓	MACs ($\times 10^9$) ↓
SAR-Opt-cGAN [8]	26.19	0.779	0.040	14.02	140.43	0.531	22.17	477.95
DSen2-CR [11]	30.29	0.878	0.034	7.008	80.37	0.431	18.95	1241.8
GLF-CR [12]	31.85	0.896	0.020	6.722	87.63	0.452	14.83	245.28
UnCRtainTS [13]	31.28	0.887	0.023	6.899	85.94	0.410	**2.02**	107.57
DiffCR [17]	32.25	0.909	0.019	5.335	71.55	0.398	22.91	137.58
DB-CR	**33.47**	**0.922**	**0.016**	**4.740**	**61.91**	**0.316**	18.06	**15.24**

network in which the input is a concatenation of the optical and SAR images.

- **GLF-CR** [12]: Global-Local Fusion Cloud Removal (GLF-CR) method utilizes advanced SAR-optical fusion to integrate both global and local feature information from SAR and optical images to estimate the cloud-free image.
- **UnCRtainTS** [13]: UnCRtainTS focuses on handling uncertainty in cloud removal tasks, using uncertainty-aware and attention-based fusion techniques to predict cloud-free images from partially cloudy observations.
- **DiffCR** [17]: DiffCR leverages a diffusion model-based approach to gradually refine cloud removal predictions, integrating a network designed with specialized blocks that fuse temporal and conditional information for enhanced performance.

For a fair comparison, we trained all baseline models with input and output sizes set to 256×256, using a learning rate of 5×10^{-5} and training for 50 epochs on the same training set as our proposed method. For DSen2-CR, we use the same network architecture setting provided by the paper. For GLF-CR, we adjusted the input and output sizes to 256×256 and conducted full-image training and inference, rather than dividing the images into patches of size of 128×128 as in the original implementation, to avoid any performance differences stemming from patch-based processing. For UnCRtainTS, the original model was designed with a lightweight purpose in mind. To prioritize performance, we scaled up the model parameter size from 0.5M to 2.02M (its performance did not continue to improve with larger sizes), which we used as the baseline. For DiffCR, we followed the original implementation and selected the Number of Function Evaluations (NFE) at inference time equal to 3, which provided the best PSNR.

Table I presents the quantitative results of all evaluated methods. We use NFE=1 for DB-CR. As shown, DB-CR consistently outperforms the others, achieving the highest PSNR and SSIM scores, and the lowest MAE and SAM values. In addition, DB-CR exhibits the best perceptual quality, reflected in the best FID and LPIPS scores. Notably, our method achieves this superior performance with a comparable number of parameters to the baseline methods, while significantly reducing MACs, which translates to faster speeds in each inference step. In Figures 5 and 6, we show some qualitative results with the RGB channels of the multispectral images visualized. Again, results obtained DB-CR are of signficantly higher quality compared to the baselines.

We further evaluate our method's performance across varying cloud cover levels. Using the S2Cloudless cloud detector [32], we generated binary masks to identify cloud pixels and calculated the percentage of cloud cover in each test image. We then computed the median quantitative metrics at 5 different cloud cover levels, ranging from 0–20% to 80–100%. As shown in Table III, our method consistently outperforms DiffCR across all cloud cover levels, demonstrating robust performance even under heavy cloud coverage.

E. Ablation study for proposed DB-CR backbone architecture

This section presents the ablation study results for the DB-CR backbone architecture. As described in Section III, we employ a two-branch U-Net architecture as the core of our model. To enhance both the speed and quality of feature extraction and improve SAR-optical image fusion, we introduced two key modifications: time-embedded NAFBlocks and SFBlocks.

As shown in Table II, the base model, which uses a two-branch U-Net with traditional residual blocks and simple concatenation ("early concatenation") for SAR-optical feature fusion, serves as the reference. Replacing the residual blocks with NAFBlocks in the backbone results in a significant improvement in reconstruction accuracy, reflected in the increased PSNR and SSIM values and the reduction in error metrics. Similarly, introducing SFBlocks ("multi-depth cross-modal attention") enhances multimodal fusion, leading to gains in performance. The combination of both NAFBlocks and SFBlocks yields the most substantial improvements, confirming their complementary roles in refining cloud removal and improving the fidelity of the final cloud-free images.

V. DISCUSSION

A. Comparison with diffusion model

Our proposed method is based on a diffusion bridge, and our results show that this is more effective at solving the cloud removal problem than a standard conditional diffusion model [17]. Traditional diffusion models employ a forward process that progressively adds noise to an image and learn a reverse process to denoise it, ultimately generating a clean version of the image. This process can be effective for image synthesis and some restoration tasks, but it often lacks control over the specific transition between distinct input and output

TABLE II
EFFECT OF BACKBONE ARCHITECTURE AND FUSION METHOD ON CLOUD REMOVAL PERFORMANCE. WE SEE THAT THE PROPOSED COMBINATION OF NAFBLOCKS IN THE BACKBONE AND ATTENTION-BASED SAR-OPTICAL FEATURE FUSION RESULTS IN THE BEST CLOUD REMOVAL PERFORMANCE.

Backbone	Fusion Method	PSNR(dB) ↑	SSIM ↑	MAE ↓	SAM ↓	FID ↓	LPIPS ↓
ResBlock	Early concatenation	30.37	0.879	0.032	6.972	79.50	0.482
NAFBlock	Early concatenation	32.75	0.916	0.018	5.190	70.11	0.368
ResBlock	Multi-depth cross-modal attention	31.95	0.894	0.024	6.251	74.69	0.437
NAFBlock	Multi-depth cross-modal attention	**33.47**	**0.922**	**0.016**	**4.740**	**61.91**	**0.316**

TABLE III
COMPARISON OF DB-CR WITH BASELINE METHOD DIFFCR (A CONDITIONAL DIFFUSION MODEL) UNDER DIFFERENT CLOUD-COVER PERCENTAGES. OUR PROPOSED DB-CR METHOD OUTPERFORMS DIFFCR FOR ALL CLOUD COVER AMOUNTS.

Cloud Cover (%)	**Method**	**PSNR ↑**	**SSIM ↑**	**MAE ↓**	**SAM ↓**	**FID ↓**	**LPIPS ↓**
0%-20%	DiffCR	35.07	0.956	0.011	2.82	56.15	0.203
	DB-CR	**36.04**	**0.964**	**0.010**	**2.71**	**44.73**	**0.150**
20%-40%	DiffCR	32.61	0.944	0.015	3.47	68.76	0.279
	DB-CR	**33.92**	**0.953**	**0.013**	**3.08**	**62.46**	**0.237**
40%-60%	DiffCR	32.48	0.926	0.015	4.02	76.88	0.324
	DB-CR	**33.55**	**0.937**	**0.013**	**3.52**	**70.54**	**0.291**
60%-80%	DiffCR	31.90	0.906	0.016	4.46	95.26	0.362
	DB-CR	**32.83**	**0.918**	**0.015**	**4.02**	**83.61**	**0.331**
80%-100%	DiffCR	30.85	0.888	0.019	5.73	123.31	0.446
	DB-CR	**31.62**	**0.903**	**0.017**	**5.09**	**106.98**	**0.407**

states, such as in the cloud-removal task where the input (cloudy) and output (cloud-free) states are well defined.

In contrast, our diffusion bridge explicitly models the transition between the cloudy image and its cloud-free counterpart, creating a targeted path that connects these two states. This allows for a more direct and constrained approach to the cloud-removal problem, ensuring that each step in the denoising process moves the image closer to a plausible cloud-free version. This characteristic provides a significant advantage over standard diffusion models, which may diverge from the desired solution due to the randomness in the reverse (denoising) process. By explicitly guiding the transition from cloudy to cloud-free states, the diffusion bridge achieves a more consistent and stable reconstruction, with reduced risk of oversmoothing or artifacts.

Moreover, the diffusion bridge formulation allows for the use of optimal transport to guide the intermediate states in a way that directly reflects the transition from cloudy to cloud-free. This more focused approach results in improved consistency and perceptual quality of the final output, reducing the risk of oversmoothing or introducing artifacts commonly seen in standard diffusion models.

Ultimately, the direct nature of the diffusion bridge makes it better suited for cloud removal, where a specific target state needs to be recovered accurately and efficiently.

B. Advantage of DB-based training for one-step inference

To further understand the benefits of our approach, we performed a comparison with a one-step end-to-end learning method that employs the same network architecture and SAR fusion as DB-CR. This ablation study method uses a single-step mapping ($T = 1$) from the cloudy image to its reference during both training and inference. In contrast, our proposed method (DB-CR) uses multiple gradual degradation levels (e.g., 1000 levels) for training but can perform inference in a single step by setting NFE = 1. Importantly, when setting NFE to 1, DB-CR essentially reduces to a single-step direct estimation from the cloudy image to its reference, achieving the same inference speed as its end-to-end learned counterpart. As demonstrated in Table IV, this single-timestep method significantly underperformed compared to our proposed strategy, which uses a more dynamic timestep size during training. We believe this discrepancy is due to the dynamic timestep acting as a form of data augmentation, making the model more robust to varying cloud conditions. By randomly sampling different timesteps during training, the model is effectively exposed to a range of intermediate states of degradation, allowing it to learn how to adapt to different cloud-cover levels. This variability enables the model to develop a more generalized understanding of the problem, akin to data augmentation, which is crucial for handling the diverse nature of cloud coverage. This adaptability ensures that the model can apply the appropriate level of reconstruction precision across different regions of an image, ultimately resulting in superior cloud-free outputs.

C. NFE-controlled distortion-perception trade-off

Previous work [19] has established that there is an inherent trade-off between image perceptual quality and distortion in reconstruction tasks, making it impossible to minimize both simultaneously. As described in Equation (12), the selection

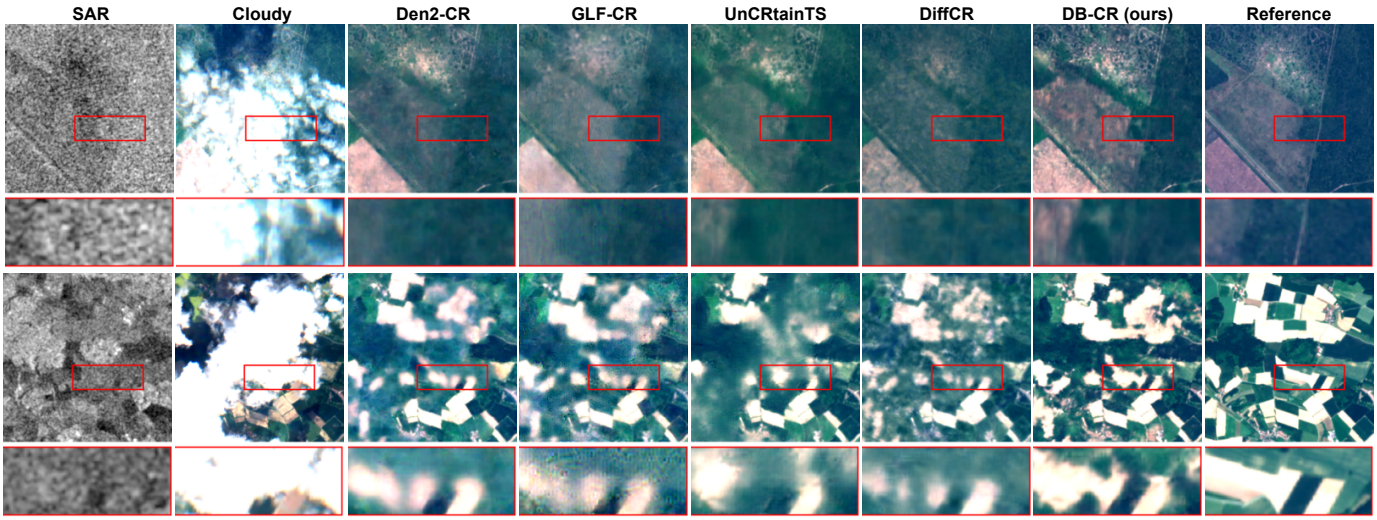

Fig. 5. Comparison of DB-CR vs. baseline methods on two agricultural scenes. For each image in the first and third rows, the detail region enclosed in the red rectangle is magnified in the subsequent row. For each scene, the original cloudy image is shown in the leftmost column, and the reference image (actual cloud-free image of the same area as the cloudy image) is in the rightmost column. Compared to the baseline methods, DB-CR achieves finer detail recovery, accurately preserves edges, and provides sharper details with superior perceptual quality.

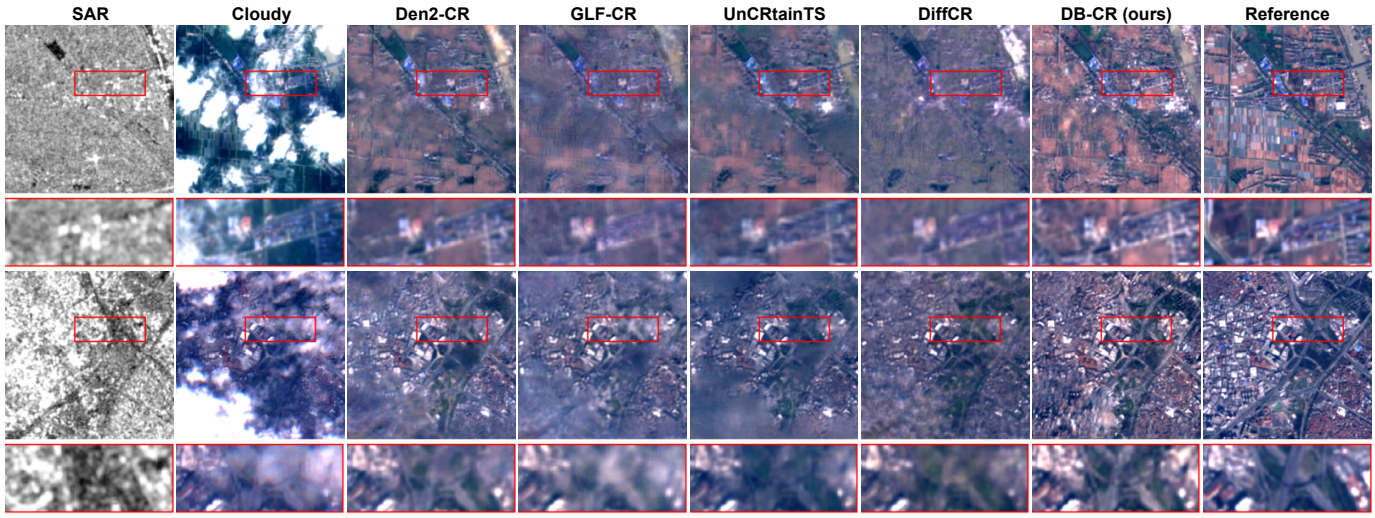

Fig. 6. Visual comparison of DB-CR vs. baseline methods on two urban scenes. DB-CR demonstrates superior detail recovery and effectively addresses the oversmoothing issues that plague all of the baseline methods.

TABLE IV
ABLATION STUDY OF DB-CR FOR DIFFUSION BRIDGE TRAINING. HERE, "W/ DB" REFERS TO OUR FULL DB-CR METHOD. IN CONTRAST, "W/O DB" REFERS TO END-TO-END TRAINING OF THE BACKBONE WITHOUT TIME EMBEDDING. ALL OTHER HYPERPARAMETERS ARE THE SAME AS FOR DB-CR

Method	PSNR ↑	SSIM ↑	MAE ↓	SAM ↓	FID ↓	LPIPS ↓
w/o DB	33.03	0.916	0.017	5.074	66.08	0.335
w/ DB	**33.47**	**0.922**	**0.016**	**4.740**	**61.91**	**0.316**

of timesteps during the inference phase introduces a degree of freedom that allows us to control the NFE (number of function evaluations for inference). According to InDI [20], different NFE values allow for a controllable distortion-perception trade-off, which is particularly significant for cloud-removal (CR) tasks. Motivated by this, we explore how the number of evaluation steps affects the multi-step restoration process in the diffusion bridge framework.

For cloud removal, reducing the NFE yields estimates that are closer to the posterior mean, effectively minimizing distortion but potentially causing over-smoothing, which can result in a loss of important details. In contrast, increasing the NFEs enhances perceptual quality, making the reconstructed images appear more visually natural, but also introduces a risk of generating unrealistic details. As shown in Table V, the choice of NFE has a direct impact on the trade-off between distortion metrics (PSNR, SSIM, MAE, SAM) and perceptual metrics (FID, LPIPS). Higher NFEs improve perceptual quality at the expense of distortion accuracy, while lower NFEs enhance distortion performance but reduce perceptual fidelity. According to Table V, setting NFE $= 3$ provides a favorable balance between PSNR and LPIPS scores, preserving cloud-

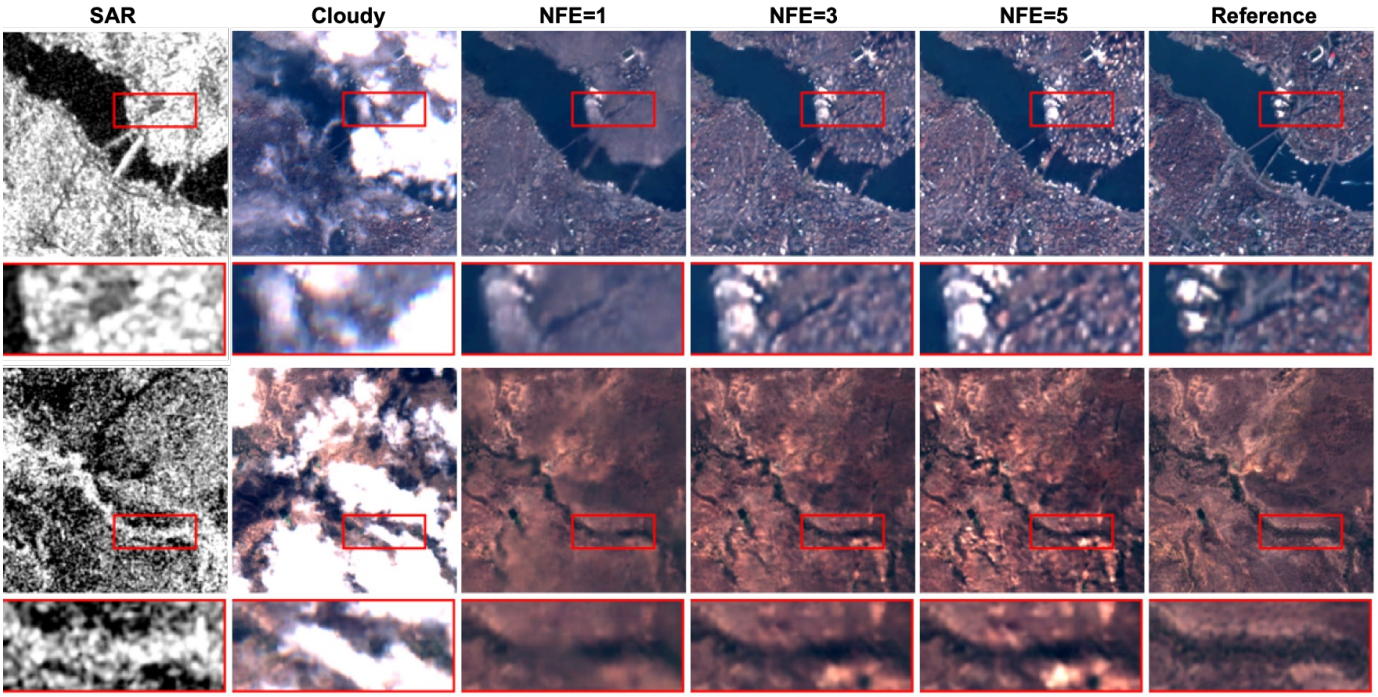

Fig. 7. Impact of NFE (number of function evaluations for inference) on DB-CR outputs. Higher NFE enhances perceptual quality by preserving finer details, while lower NFE provides smoother prediction, balancing visual clarity and data consistency.

free accuracy while maintaining good perceptual quality. Furthermore, observing results in Table I, running DB-CR with both NFE $= 1$ and NFE $= 3$ configurations outperform all baseline methods across distortion and perceptual metrics.

For applications requiring higher visual fidelity, increasing NFE can further improve perceptual quality, albeit at the cost of increased distortion. For instance, as illustrated in the top row of Figure 7, higher NFE results in more distinct gaps appearing in the terrain rather than smooth connections. Similarly, in the bottom row, higher NFE makes it increasingly clear that the area beneath the cloud contains buildings, as opposed to the blurry artifacts seen with NFE $= 1$.

D. Influence of stochastic perturbation for CR

We investigate the impact of introducing stochastic perturbation into the problem formulation. In general image restoration tasks, the effect of stochastic perturbation tends to be task-dependent. For instance, in JPEG restoration, stochastic perturbation has been shown to enhance performance, whereas in deblurring, it often leads to degradation in quality [20], [21]. Therefore, it is valuable to evaluate the influence of stochastic perturbation on the cloud removal task.

In our formulation, stochastic perturbation can effectively convert the ODE-based formulation of DB-CR into an SDE-based approach. Specifically, we modify the forward process of the diffusion bridge as follows:

$$\boldsymbol{x}_t = (1-\alpha_t)\boldsymbol{x}_0 + \alpha_t \boldsymbol{y} + \beta_t \epsilon, \quad \alpha_t \in [0,1], \beta_t \in [0,1],$$

where $\beta_t \epsilon_t$ is the standard Brownian motion having zero mean and covariance $\beta_t \mathbf{I}$ at time t. In this SDE-based approach, the DB network is also trained to perform MMAE estimation ($\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_t]$) of the target cloud-free image using the same training strategy as the ODE-case in Equation (11). For inference, however the SDE case will be different from the ODE case shown in Equation (12). Instead, inference in the SDE case uses:

$$\begin{aligned}\mathbb{E}[\boldsymbol{x}_{t-1}|\boldsymbol{x}_t] = \left(1-\tfrac{\alpha_{t-1}}{\alpha_t}\right)\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_t] + \left(\tfrac{\alpha_{t-1}}{\alpha_t}\right)\boldsymbol{x}_t \\ + \left(\beta_t \tfrac{\alpha_{t-1}}{\alpha_t} - \beta_{t-1}\right)\epsilon.\end{aligned} \tag{17}$$

TABLE V
EFFECT OF NUMBER OF FUNCTION EVALUATIONS (NFE) THROUGH THE DB-CR BACKBONE ON CLOUD REMOVAL PERFORMANCE. DISTORTION METRICS ARE BEST WITH NFE= 1, WHILE PERCEPTUAL METRICS IMPROVE WITH HIGHER NFE. THIS CONFIRMS THE WELL-KNOWN PERCEPTION-DISTORTION TRADE-OFF IN IMAGE RESTORATION.

NFE	PSNR ↑	SSIM ↑	MAE ↓	SAM ↓	FID ↓	LPIPS ↓
1	**33.47**	**0.922**	**0.016**	**4.740**	61.91	0.316
3	33.10	0.915	0.017	5.031	**58.03**	0.273
5	32.51	0.905	0.018	5.288	58.40	**0.269**

TABLE VI
PERFORMANCE COMPARISON OF DB-CR WITH SDE AND ODE FORMULATION.

Method	PSNR ↑	SSIM ↑	MAE ↓	SAM ↓	FID ↓	LPIPS ↓
DB-CR (SDE)	32.95	0.914	0.018	5.133	68.51	0.369
DB-CR (ODE)	**33.47**	**0.922**	**0.016**	**4.740**	**61.91**	**0.316**

As shown in Table VI, our experiments indicate that the SDE-based approach is not well-suited for the cloud removal problem, leading to reduced performance in both distortion

and perception metrics. The added randomness, while theoretically capable of preventing over-smoothing and encouraging diverse solutions, often results in inconsistent output quality. In practice, the cloud removal task benefits more from a deterministic approach where the reconstruction process is guided precisely to recover missing details without introducing additional uncertainty. The iterative refinement in the ODE-based method ensures a more controlled transition from the cloudy to the clean state, leading to better preservation of fine details and overall perceptual quality.

E. Influence of inference step size

As shown in Algorithm 2, the default inference step size is uniformly set as $s = \frac{T}{N}$, corresponding to evenly spaced time intervals. In practice, however, the step size s can be non-uniform. For instance, the steps may be smaller at the beginning and larger toward the end of the sampling process, or the opposite. We refer to the case where smaller steps are concentrated near T as *biased toward* T, and the case where smaller steps are concentrated near 0 as *biased toward 0*.

To implement such biases, we define a monotonically increasing sequence $\{\tau_n\}_{n=0}^{N}$ with $\tau_0 = 0$ and $\tau_N = T$. One simple way is to use a nonlinear warping of the uniform grid:

$$\tau_n = T \cdot \left(\frac{n}{N}\right)^{\gamma}, \quad n = 0, 1, \ldots, N \tag{18}$$

where $\gamma > 0$ is a control parameter:

- $\gamma = 1$: corresponds to the uniform step size we used in the experiments reported above.
- $\gamma > 1$: yields steps biased toward $t = 0$.
- $\gamma < 1$: yields steps biased toward $t = T$.

The step sizes are then computed as $s_n = \tau_{n+1} - \tau_n$ for $n = 0, \ldots, N-1$. As shown in Table VII, biasing toward early steps improves distortion but harms perception, while biasing toward later steps does the opposite. Uniform steps offer a balanced trade-off.

TABLE VII
EFFECT OF SAMPLING STEP SIZE IN THE SAMPLING PRECESSING IN CASE OF NFE=3.

γ	PSNR ↑	SSIM ↑	MAE ↓	SAM ↓	FID ↓	LPIPS ↓
0.2	33.01	0.912	0.017	5.035	**57.95**	0.274
1	33.10	0.915	0.017	5.031	58.03	**0.273**
2	**33.19**	**0.918**	**0.016**	**5.017**	58.31	0.277

VI. LIMITATIONS & FUTURE WORK

One limitation of the current study pertains to the handling of speckle noise, which is a common artifact in SAR imagery. Our approach did not incorporate explicit SAR denoising preprocessing steps; rather, we followed the protocol of earlier methods, e.g., [11], [17] and depend on the capacity of our multimodal architecture to implicitly discern underlying features from the noisy SAR input. One direction for future work will be to assess if the integration of SAR denoising algorithms as a preliminary step can yield further improvements in the efficacy and robustness of the cloud removal DB. Another interesting direction would be to explore the structural consistency [33], [34] in SAR and optical images as additional control to further improve the reconstruction performance.

Additionally, the forward process within our diffusion bridge framework was intentionally simplified to linear interpolation. This design choice was made for a straightforward interpretation of performance, particularly in relation to the MMAE estimator and the established perception-distortion trade-off. In future work, a promising research direction involves developing methodologies to learn both the forward and reverse diffusion processes. Employing a dual-branch DB architecture to learn these processes dynamically could enable a more nuanced and accurate representation of the cloud corruption phenomenon itself, potentially leading to superior restoration outcomes.

VII. CONCLUSION

In this paper, we introduced DB-CR, a novel diffusion bridge-based method designed for cloud removal. DB-CR leverages fixed start and end conditions to guide a precise, step-by-step reduction of cloud cover, closely aligning with the desired cloud-free state. By applying optimal transport principles, DB-CR achieves an optimal balance of fidelity and efficiency at each intermediate step, preserving structural integrity and perceptual quality throughout the process. Our diffusion bridge-based network backbone, coupled with the fusion of features from both the SAR and optical modalities, establishes DB-CR as the leading approach for both accuracy and computational efficiency. Experimental results validate DB-CR's state-of-the-art performance in terms of both distortion metrics and perceptual quality.

REFERENCES

[1] F. Stumpf, M. K. Schneider, A. Keller, A. Mayr, T. Rentschler, R. G. Meuli, M. Schaepman, and F. Liebisch, "Spatial monitoring of grassland management using multi-temporal satellite imagery," *ECOL INDIC*, vol. 113, p. 106201, 2020.

[2] J. Yin, J. Dong, N. A. Hamm, Z. Li, J. Wang, H. Xing, and P. Fu, "Integrating remote sensing and geospatial big data for urban land use mapping: A review," *International Journal of Applied Earth Observation and Geoinformation*, vol. 103, p. 102514, 2021.

[3] M. D. King, S. Platnick, W. P. Menzel, S. A. Ackerman, and P. A. Hubanks, "Spatial and temporal distribution of clouds observed by modis onboard the terra and aqua satellites," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 7, pp. 3826–3852, 2013.

[4] T. F. Chan and J. Shen, "Nontexture inpainting by curvature-driven diffusions," *Journal of Visual Communication and Image Representation*, vol. 12, no. 4, pp. 436–449, 2001.

[5] G.-L. Huang and P.-Y. Wu, "CTGAN: Cloud transformer generative adversarial network," in *IEEE Conference on Image Processing*, 2022, pp. 511–515.

[6] K. Enomoto, K. Sakurada, W. Wang, H. Fukui, M. Matsuoka, R. Nakamura, and N. Kawaguchi, "Filmy cloud removal on satellite imagery with multispectral conditional generative adversarial nets," in *Computer Vision and Pattern Recognition Workshops*, 2017, pp. 48–56.

[7] H. Pan, "Cloud removal for remote sensing imagery via spatial attention generative adversarial network," *arXiv preprint arXiv:2009.13015*, 2020.

[8] C. Grohnfeldt, M. Schmitt, and X. Zhu, "A conditional generative adversarial network to fuse SAR and multispectral optical data for cloud removal from sentinel-2 images," in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, 2018, pp. 1726–1729.

[9] J. Bermudez, P. Happ, D. Oliveira, and R. Feitosa, "SAR to optical image synthesis for cloud removal with generative adversarial networks," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 4, pp. 5–11, 2018.

[10] V. Sarukkai, A. Jain, B. Uzkent, and S. Ermon, "Cloud removal from satellite images using spatiotemporal generator networks," in *Winter Conference on Applications in Computer Vision*, 2020, pp. 1796–1805.

[11] A. Meraner, P. Ebel, X. X. Zhu, and M. Schmitt, "Cloud removal in sentinel-2 imagery using a deep residual neural network and SAR-optical data fusion," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 166, pp. 333 – 346, 2020.

[12] F. Xu, Y. Shi, P. Ebel, L. Yu, G.-S. Xia, W. Yang, and X. X. Zhu, "GLF-CR: SAR-enhanced cloud removal with global–local fusion," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 192, pp. 268–278, 2022.

[13] P. Ebel, V. S. F. Garnot, M. Schmitt, J. Wegner, and X. X. Zhu, "Uncrtaints: Uncertainty quantification for cloud removal in optical satellite time series," in *Computer Vision and Pattern Recognition Workshops*, 2023.

[14] R. Bamler, "Principles of synthetic aperture radar," *Surveys in Geophysics*, vol. 21, no. 2, pp. 147–157, 2000.

[15] R. Jing, F. Duan, F. Lu, M. Zhang, and W. Zhao, "Denoising diffusion probabilistic feature-based network for cloud removal in sentinel-2 imagery," *Remote Sensing*, vol. 15, no. 9, 2023.

[16] X. Zhao and K. Jia, "Cloud removal in remote sensing using sequential-based diffusion models," *Remote Sensing*, vol. 15, no. 11, p. 2861, 2023.

[17] X. Zou, K. Li, J. Xing, Y. Zhang, S. Wang, L. Jin, and P. Tao, "DiffCR: A fast conditional diffusion framework for cloud removal from optical satellite images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–14, 2024.

[18] J. Sui, Y. Ma, W. Yang, X. Zhang, M.-O. Pun, and J. Liu, "Diffusion enhancement for cloud removal in ultra-resolution remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[19] Y. Blau and T. Michaeli, "The perception-distortion tradeoff," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6228–6237.

[20] M. Delbracio and P. Milanfar, "Inversion by direct iteration: An alternative to denoising diffusion for image restoration," *Transactions on Machine Learning Research*, 2023.

[21] G. Liu, A. Vahdat, D. Huang, E. A. Theodorou, W. Nie, and A. Anandkumar, "I2SB: image-to-image schrödinger bridge," in *International Conference on Machine Learning (ICML)*, 2023, pp. 22 042–22 062.

[22] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[23] Y. Wang, J. Yu, and J. Zhang, "Zero-shot image restoration using denoising diffusion null-space model," *The Eleventh International Conference on Learning Representations*, 2023.

[24] R. Jing, F. Duan, F. Lu, M. Zhang, and W. Zhao, "Denoising diffusion probabilistic feature-based network for cloud removal in sentinel-2 imagery," *Remote Sensing*, vol. 15, no. 9, p. 2217, 2023.

[25] Z. Luo, F. K. Gustafsson, Z. Zhao, J. Sjölund, and T. B. Schön, "Refusion: Enabling large-size realistic image restoration with latent-space diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 1680–1691.

[26] Y. Hu, A. Peng, W. Gan, P. Milanfar, M. Delbracio, and U. S. Kamilov, "Stochastic deep restoration priors for imaging inverse problems," *arXiv preprint arXiv:2410.02057*, 2024.

[27] L. Chen, X. Chu, X. Zhang, and J. Sun, "Simple baselines for image restoration," in *European Conference on Computer Vision*, 2022, pp. 17–33.

[28] K. Chen and L. Yu, "Motion deblur by learning residual from events," *IEEE Transactions on Multimedia*, 2024.

[29] L. Sun, C. Sakaridis, J. Liang, Q. Jiang, K. Yang, P. Sun, Y. Ye, K. Wang, and L. V. Gool, "Event-based fusion for motion deblurring with cross-modal attention," in *European conference on computer vision*. Springer, 2022, pp. 412–428.

[30] M. Schmitt, L. H. Hughes, C. Qiu, and X. X. Zhu, "SEN12MS–a curated dataset of georeferenced multi-spectral sentinel-1/2 imagery for deep learning and data fusion," *arXiv preprint arXiv:1906.07789*, 2019.

[31] A. Meraner, P. Ebel, X. X. Zhu, and M. Schmitt, "Cloud removal in sentinel-2 imagery using a deep residual neural network and SAR-optical data fusion," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 166, pp. 333–346, 2020.

[32] A. Zupanc, "Improving cloud detection with machine learning," https://medium.com/sentinel-hub/improving-cloud-detection-with-machine-learning-c09dc5d7cf13, 2019, available online: (accessed on 26 January 2025).

[33] Y. Sun, L. Lei, X. Li, X. Tan, and G. Kuang, "Structure consistency-based graph for unsupervised change detection with homogeneous and heterogeneous remote sensing images," *IEEE transactions on geoscience and remote sensing*, vol. 60, pp. 1–21, 2021.

[34] Y. Sun, L. Lei, D. Guan, G. Kuang, Z. Li, and L. Liu, "Locality preservation for unsupervised multimodal change detection in remote sensing imagery," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 4, pp. 6955–6969, 2024.