remote sensing

MDPI

Article

SAR-Conditioned Consistency Model for Effective Cloud Removal in Remote Sensing Images

Qizhuo Han [1,†], Bo Huang [2,†] and Ying Li [1,*]

[1] School of Computer Science, Northwestern Polytechnical University, Xi'an 710129, China; hqz@mail.nwpu.edu.cn
[2] School of Software, Northwestern Polytechnical University, Xi'an 710129, China; bohuang@mail.nwpu.edu.cn
* Correspondence: lybyp@nwpu.edu.cn
† These authors contributed equally to this work.

Highlights

What are the main findings?

- We propose CM-CR, a SAR-conditioned cloud-removal framework that distills a SAR-Conditioned Score-Based Diffusion (teacher) into a SAR-conditioned consistency model (student) and refines outputs via multistep resampling.
- CM-CR achieves state-of-the-art PSNR/SSIM/SAM/MAE on SEN12MS-CR while requiring substantially fewer sampling steps than standard diffusion models.

What is the implications of the main findings?

- The method improves the usability of optical remote sensing data under thick-cloud conditions across diverse scenes.
- It preserves the reconstruction fidelity of diffusion models with markedly reduced inference complexity, making it suitable for large-scale and operational remote sensing tasks.

Abstract

Cloud contamination, especially thick cloud cover, severely limits the usability of optical remote sensing imagery by obscuring surface information. Due to the strong penetrability of microwave signals, Synthetic Aperture Radar (SAR) has emerged as an effective source for thick cloud removal. While SAR-assisted deep learning methods, such as CNNs and GANs, have made notable progress, the quality of generated imagery still requires improvement. Diffusion models, which offer strong potential for enhancing generation fidelity, could address this limitation but suffer from slow sampling speeds that constrain practical use and underscore the need for greater efficiency. To simultaneously enhance both reconstruction quality and sampling efficiency, this paper proposes a fast-sampling SAR-conditioned consistency model based on consistency distillation, named CM-CR, which adopts a teacher–student architecture to divide the reconstruction process into a rapid coarse prediction stage and a detailed refinement stage, significantly reducing per-scene processing time while maintaining high reconstruction fidelity. Specifically, a SAR-Conditioned Score-Based Diffusion Model (SCSBD) is first developed as the teacher network for learning a SAR-conditioned optical image generation model. Consistency distillation is then used to derive the student network SAR-conditioned consistency model (SCCM), which enables a rapid coarse prediction through single-step sampling. Finally, a Progressive Denoising via Multistep Resampling (PDMSR) strategy is introduced to iteratively refine the single-step output, producing fine-grained reconstructions. Comparative experiments conducted on the widely used cloud removal benchmark dataset

check for updates

Academic Editor: Xinghua Li

Received: 26 August 2025
Revised: 6 November 2025
Accepted: 12 November 2025
Published: 14 November 2025

Citation: Han, Q.; Huang, B.; Li, Y. SAR-Conditioned Consistency Model for Effective Cloud Removal in Remote Sensing Images. *Remote Sens.* **2025**, *17*, 3721. https://doi.org/10.3390/rs17223721

Copyright: © 2025 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) license (https://creativecommons.org/licenses/by/4.0/).

SEN12MS-CR demonstrate that the proposed CM-CR method achieves state-of-the-art (SOTA) performance across all image quality metrics. Notably, although its design uses approximately 80 times more parameters compared with a standard Denoising Diffusion Probabilistic Model (DDPM), it delivers up to a 40-fold acceleration at inference.

Keywords: cloud removal; SAR; optical remote sensing images; consistency model

1. Introduction

With the rapid advancement of remote sensing technology, optical remote sensing imagery has become indispensable for a wide range of applications, including meteorological monitoring [1], agricultural assessment [2], urban planning [3], and disaster management [4]. High-resolution and wide-coverage optical images offer crucial support for obtaining surface information on Earth. However, these images are frequently compromised by atmospheric interference such as clouds and haze, which significantly degrade image quality and usability. Among these factors, cloud contamination is the most prominent, as it not only obscures surface features and causes information loss, but also introduces artifacts and blurring, thereby hindering subsequent analysis and interpretation [5]. Effectively removing cloud interference from optical remote sensing images has thus emerged as a critical challenge in the field.

Cloud cover can be categorized into thin and thick clouds based on their optical thickness and impact on surface visibility. Thin clouds partially obscure ground features, whereas thick clouds block surface information entirely, often causing severe spectral distortion and shadow effects [6]. The task of thick cloud removal aims to recover ground information from regions severely occluded by dense clouds. Existing methods for thick cloud removal can be broadly grouped into three categories: inpainting-based, multi-temporal, and multi-modal approaches [7].

Inpainting-based methods assume spatial and semantic correlations between cloud-covered and cloud-free regions within the same image [8–13]. These methods use the surrounding cloud-free areas to infer and reconstruct the missing content. While they can produce visually plausible results when the scene is simple and cloud coverage is limited, their reliance on single-modal information and lack of prior knowledge limits their ability to recover complex surface features accurately.

Multi-temporal methods [14–20] leverage multiple images of the same region acquired at different times, using cloud-free references to reconstruct cloud-obscured regions. However, these approaches often suffer from difficulties in acquiring temporally aligned and high-quality references, and inconsistencies caused by temporal differences may degrade reconstruction performance.

Multi-modal approaches [21–27] address these limitations by incorporating complementary data from different sensors. In particular, Synthetic Aperture Radar (SAR), with its all-weather and day-and-night imaging capabilities, can provide reliable structural and textural information for cloud-covered areas. By circumventing the temporal inconsistencies associated with multi-temporal methods, SAR data offer a promising modality for the accurate reconstruction of heavily clouded regions. Accordingly, this work adopts SAR imagery as a conditioning input to guide the cloud removal process.

Generative Adversarial Networks (GANs) have been widely applied in deep learning-based cloud removal frameworks [28–30]. While effective, GANs often suffer from inherent instability during training, difficulties in convergence, and vulnerability to mode collapse. Moreover, the limited generative capacity and poor generalization of GAN-based methods

frequently result in blurred textures and loss of fine details, which restrict their ability to meet the precision requirements for high-quality cloud removal.

Recently, diffusion models [31] have emerged as state-of-the-art generative models, outperforming GANs in various vision tasks [32]. Built upon a Markovian denoising process, diffusion models offer more stable and reliable training along with superior generative fidelity. They have been successfully applied to a wide range of image generation tasks such as inpainting [33], super-resolution [34], and style transfer [35]. These advancements have inspired the use of diffusion models in cloud removal, with several works reporting promising results [36–41]. However, vanilla diffusion-based approaches still suffer from slow inference, as the iterative denoising process required for high-quality synthesis incurs high computational costs. This inefficiency remains a major bottleneck for the practical deployment of diffusion-based cloud removal techniques.

To address the inefficiency of diffusion models, recent advances have introduced consistency models [42], which preserve high generative fidelity and accelerate inference, providing an efficient solution for cloud removal tasks. Inspired by this line of work, we propose CM-CR, a novel and efficient cloud removal framework based on SAR-conditioned consistency models. It is designed to simultaneously achieve high-quality reconstruction and fast inference, enabling practical cloud removal for optical remote sensing images. The main contributions of this work are as follows:

1. We develop a SAR-Conditioned Score-Based Diffusion (SCSBD) model as a teacher network, where SAR data are utilized as conditional input to guide the generation of structural features and spatial layouts under SAR constraints, establishing a solid foundation for the subsequent distillation process.
2. We introduce a consistency distillation strategy that distills the SCCM (SAR-Conditioned Consistency Model) student model from the trained SCSBD, enabling efficient single-step cloud removal with coarse predictions.
3. To further refine the initial output, we propose a Progressive Denoising via Multi-step Resampling (PDMSR) strategy. By injecting noise into the coarse results and resampling, this strategy recovers fine-grained spectral details, effectively balancing inference efficiency and reconstruction quality.

2. Preliminaries

2.1. Diffusion Models

Diffusion models are a family of generative models that formulate data synthesis as a two-stage process. The forward process gradually perturbs data with Gaussian noise, while the reverse process, parameterized by a neural network, iteratively reconstructs the original distribution. The seminal denoising diffusion probabilistic model (DDPM) [32] first established this formulation, and Song et al. [43] further extended it through a stochastic differential equation (SDE) framework, leading to score-based generative models capable of accurately approximating complex data distributions. Our method builds upon the score-based diffusion models, and the following section provides a concise theoretical overview.

In the SDE-based formulation of diffusion models, data $\mathbf{x}_0 \sim p_{\text{data}}(\mathbf{x})$ are progressively perturbed along a continuous time variable $t \in [0, T]$. The forward process is defined by the following equation:

$$d\mathbf{x}_t = \mu(\mathbf{x}_t, t)\, dt + \sigma(t)\, d\mathbf{w}_t, \tag{1}$$

where $\mu(\mathbf{x}_t, t)$ is the drift term, $\sigma(t)$ controls the noise intensity, and $\mathbf{w}_t$ denotes a standard Wiener process. Solving this SDE transforms the data distribution p_{data} into a simple prior, typically a standard Gaussian $p_T(\mathbf{x})$, which serves as the starting point for the generative

reverse process. To generate new data, sampling begins by drawing noise from the prior $\hat{\mathbf{x}}_T \sim \mathcal{N}(0, T^2\mathbf{I})$ and reversing the diffusion process. The reverse-time SDE can be defined as:

$$d\mathbf{x} = \Big[\mu(\mathbf{x}, t) - \sigma(t)^2 \nabla_{\mathbf{x}} \log p_t(\mathbf{x})\Big] dt + \sigma(t)\, d\bar{\mathbf{w}}, \tag{2}$$

where $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$ is the score function and $\bar{\mathbf{w}}$ denotes a Wiener process with time flowing backward from T to 0. By integrating this reverse SDE, samples $\hat{\mathbf{x}}_0$ can be obtained, which approximate the target distribution $p_{\text{data}}(\mathbf{x})$.

Song et al. [43] demonstrated that the reverse SDE has an equivalent deterministic formulation, termed the probability flow ODE (PF-ODE), which preserves the same marginal distributions:

$$d\mathbf{x} = \Big[\mu(\mathbf{x}, t) - \tfrac{1}{2}\sigma(t)^2 \nabla_{\mathbf{x}} \log p_t(\mathbf{x})\Big] dt. \tag{3}$$

In practice, a score network $\mathbf{s}_\phi(\mathbf{x}, t) \approx \nabla_{\mathbf{x}} \log p_t(\mathbf{x})$ is trained via score matching. Under the setting of Karras et al. [44], where $\mu(\mathbf{x}, t) = 0$ and $\sigma(t) = \sqrt{2t}$, Equation (3) reduces to

$$\frac{d\mathbf{x}_t}{dt} = -t\, \mathbf{s}_\phi(\mathbf{x}_t, t). \tag{4}$$

2.2. Consistency Models

Consistency models [42] have recently been proposed to support efficient single-step sampling while still permitting iterative generation for improved trade-offs between sample quality and computational cost. Unlike conventional diffusion models that rely on lengthy sampling trajectories, consistency models directly learn a consistency function $f : (\mathbf{x}_t, t) \mapsto \mathbf{x}_\epsilon$ that enforces self-consistency across different time steps along the PF-ODE trajectory, $f(\mathbf{x}_t, t) = f(\mathbf{x}_{t'}, t')$ for all $t, t' \in [\epsilon, T]$, where ϵ is a fixed small positive number. This property enables them to achieve one-step generation.

To ensure reliable training, a boundary condition $f(\mathbf{x}_\epsilon, \epsilon) = \mathbf{x}_\epsilon$ is imposed, guaranteeing identity mapping at the start of the trajectory. In practice, the consistency model $f_\theta(\mathbf{x}, t)$ is parameterized by a neural network subject to this constraint:

$$f_\theta(\mathbf{x}, t) = c_{\text{skip}}(t)\mathbf{x} + c_{\text{out}}(t) F_\theta(\mathbf{x}, t) \tag{5}$$

where $c_{\text{skip}}(t)$ and $c_{\text{out}}(t)$ are differentiable functions with the $c_{\text{skip}}(\epsilon) = 1$ and $c_{\text{out}}(\epsilon) = 0$. $F_\theta(\mathbf{x}, t)$ denotes the output of the deep neural network. Sampling with a trained consistency model is highly efficient: starting from Gaussian noise $\hat{\mathbf{x}}_T \sim \mathcal{N}(0, T^2 I)$, a single forward evaluation $\hat{\mathbf{x}}_\epsilon = f_\theta(\hat{\mathbf{x}}_T, T)$ suffices to produce high-quality samples. These properties distinguish consistency models as an independent family of generative models, complementing and extending diffusion-based approaches.

3. Methodology

We propose CM-CR, a cloud removal framework built upon SAR-conditioned consistency modeling. As illustrated in Figure 1, CM-CR comprises three main components.

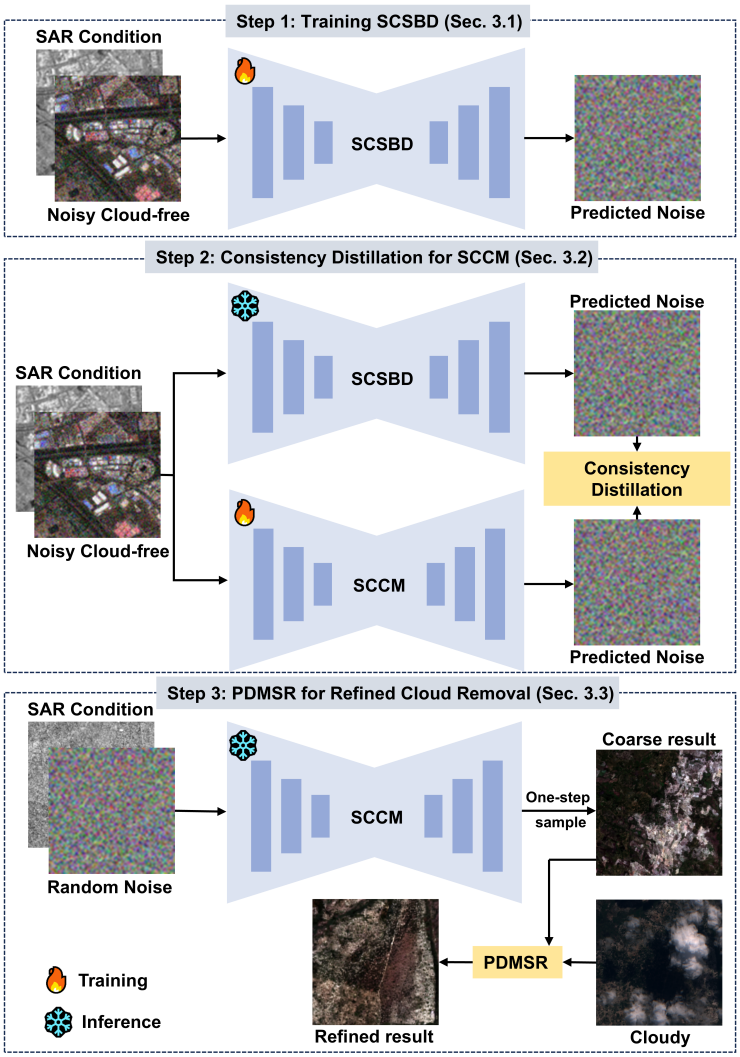


Figure 1. Overall architecture of the proposed CM-CR framework. It consists of three main components: (1) training the SCSBD model as the teacher; (2) consistency distillation for student model SCCM; and (3) refining the coarse output via the PDMSR strategy to obtain high-quality cloud-free images. Step 1 trains the SCSBD as the teacher model, as elaborated in Section 3.1. Step 2 employs consistency distillation to transfer knowledge from SCSBD to the student model SCCM, enabling single-step sampling for coarse cloud removal, detailed in Section 3.2. Step 3 further refines these coarse results via the proposed PDMSR strategy, yielding high-quality, cloud-free reconstructions, as described in Section 3.3.

3.1. SAR-Conditioned Score-Based Diffusion (SCSBD) Model

As mentioned in Section 2.2, we need to train a teacher network to estimate the score, which will be used for subsequent sampling and cloud removal processes. Figure 2 illustrates the training procedure of our SCSBD network. We utilize U-Net [45] as the backbone network to construct the SCSBD network. The U-Net architecture, with its symmetrical encoder–decoder structure, iteratively compresses and reconstructs the spatial features of the image, while employing skip connections to retain high-resolution spatial information. Owing to its all-weather and all-day imaging capability, SAR imagery provides stable ground information even under cloud coverage. This feature enables SAR data to effectively compensate for the missing information in optical images caused by cloud occlusion, significantly enhancing the model's ability to reconstruct realistic scenes.

To further improve the accuracy of cloud removal, we integrate SAR imagery as auxiliary conditional information, which is concatenated with noisy cloud-free optical images to serve as input to train the SCSBD. Specifically, the SAR image $\mathbf{c}_{sar}$ is merged with the noisy optical

image $\mathbf{x}_t$ at time step t and jointly processed as input to the model. This fusion approach ensures effective integration of SAR and optical features at the input stage, allowing the model to effectively capture their relationships during encoding, thereby leveraging SAR data as prior knowledge to enhance scene understanding and reconstruction.

Within the U-Net architecture, SAR data is not only combined with cloud-free optical images at the initial input stage but also further incorporated via skip connections during the decoder phase to facilitate better learning and utilization of SAR conditional information. Ultimately, the model produces a predicted score $\mathbf{s}\phi(\mathbf{x}, \sigma, \mathbf{c}_{sar})$ under the SAR condition, which guides the reconstruction process. This approach allows the model to manage complex weather conditions and occlusions more effectively, generating clearer and more accurate cloud-free images.

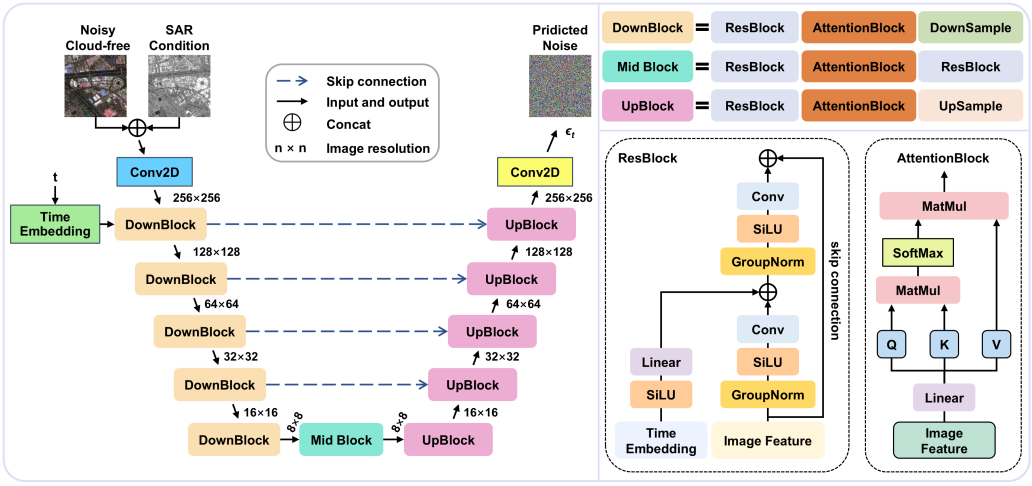


Figure 2. Architecture of the proposed SCSBD and SCCM. The network adopts a U-Net-like encoder–decoder structure to iteratively downsample and upsample the noisy optical image, conditioned on the corresponding SAR image and time embedding. The final predicted noise is estimated to guide the sampling process for effective cloud removal.

In the following sections, we will elaborate on the modeling details of the SCSBD from a theoretical perspective. We define $p_\sigma(\tilde{\mathbf{x}}) = \int p_{\text{data}}(\mathbf{x})\mathcal{N}(\tilde{\mathbf{x}}; \mathbf{x}, \sigma^2 I)dt$ as the perturbed data distribution. We denote $\{\sigma_i\}_{i=1}^{L}$ to be a positive noise levels sequence that satisfies $\sigma_{min} = \sigma_1 < \sigma_2 < \cdots < \sigma_L = \sigma_{\max}$. The noise levels are selected such that σ_1 is small enough to minimize its impact on the data $p_{\sigma_{\min}}(\mathbf{x}) \approx p_{\text{data}}(\mathbf{x})$, while σ_L is large enough such that $p_{\sigma_{\max}}(\mathbf{x}) \approx \mathcal{N}(\mathbf{x}; \mathbf{0}, \sigma_{\max}^2 \mathbf{I})$. Our goal is to train an SAR-conditioned score network that can estimate the scores of all perturbed data distributions simultaneously. Specifically, for each $\sigma \in \{\sigma_i\}_{i=1}^{L}$, the relationship $\mathbf{s}\phi(\tilde{\mathbf{x}}, \sigma, \mathbf{c}_{sar}) \approx \nabla_{\tilde{\mathbf{x}}} \log p_\sigma(\tilde{\mathbf{x}}|\mathbf{c}_{sar})$ should hold. Note that $\mathbf{s}\phi(\mathbf{x}, \sigma, \mathbf{c}_{sar}) \in \mathbb{R}^D$ when $\mathbf{x} \in \mathbb{R}^D$. We refer to $\mathbf{s}\phi(\mathbf{x}, \sigma, \mathbf{c}_{sar})$ as the SCSBD.

The denoising score matching method was employed to train the SCSBD. As shown in Algorithm 1, let the noise distribution be $p_\sigma(\tilde{\mathbf{x}} \mid \mathbf{x}) = \mathcal{N}(\tilde{\mathbf{x}} \mid \mathbf{x}, \sigma^2 I)$. Consequently, the gradient of the log-probability is given by $\nabla_{\tilde{\mathbf{x}}} \log p_\sigma(\tilde{\mathbf{x}} \mid \mathbf{x}) = -\frac{\tilde{\mathbf{x}}-\mathbf{x}}{\sigma^2}$. For a specific value of σ, the objective function for denoising score matching is:

$$\ell(\phi; \sigma_i) = \frac{1}{2}\mathbb{E}_{p_{\text{data}}(\mathbf{x})}\mathbb{E}_{\tilde{\mathbf{x}}_i \sim \mathcal{N}(\mathbf{x}_i, \sigma_i^2 I)}\left[\left\| s_\phi(\tilde{\mathbf{x}}_i, \sigma_i, \mathbf{c}_{sar}) + \frac{\tilde{\mathbf{x}}_i - \mathbf{x}_i}{\sigma_i^2} \right\|_2^2\right] \tag{6}$$

By performing a weighted summation of all $\sigma \in \{\sigma_i\}_{i=1}^{L}$ objectives, we obtain a final objective:

$$\mathcal{L}(\phi;\{\sigma_i\}_{i=1}^{L}) = \frac{1}{L}\sum_{i=1}^{L}\lambda(\sigma_i)\ell(\phi;\sigma_i) \tag{7}$$

where $\lambda(\sigma_i) > 0$ be a coefficient function dependent on σ_i. Provided that $\mathbf{s}_\phi(\mathbf{x},\sigma,\mathbf{c}_{sar})$ is well trained, the function $\mathbf{s}_{\phi^*}(\mathbf{x},\sigma,\mathbf{c}_{sar})$ will minimize Equation (7) for almost all $i \in \{1,2,\cdots,L\}$.

Algorithm 1 Training SCSBD via Denoising Score Matching

Require: Training dataset $D = \{\mathbf{x}_i\}_{i=1}^{N}$, noise levels $\{\sigma_i\}_{i=1}^{L}$, coefficients $\lambda(\sigma)$, Model $\mathbf{s}_\phi(\mathbf{x},\sigma,\mathbf{c}_{sar})$, learning rate η, number of epochs E, batch size B
Ensure: Trained model parameters ϕ

for epoch = 1 to E **do**
  for each data point x_i in D **do**
    for each noise level σ_i **do**
      Sample noise $\epsilon \sim \mathcal{N}(0, I)$
      Perturb data $\tilde{\mathbf{x}}_i \leftarrow \mathbf{x}_i + \sigma_i\epsilon$
      Noise distribution $p_{\sigma_i}(\tilde{\mathbf{x}}_i|\mathbf{x}_i) = \mathcal{N}(\tilde{\mathbf{x}}_i|\mathbf{x}_i,\sigma_i^2 I)$
      Compute the score of the perturbed data $\nabla_{\tilde{\mathbf{x}}_i}\log p_{\sigma i}(\tilde{\mathbf{x}}_i \mid \mathbf{x}_i) = -(\tilde{\mathbf{x}}_i - \mathbf{x}_i)/\sigma_i^2$
      Estimate the score by model $\mathbf{s}_\phi(\tilde{\mathbf{x}}_i,\sigma_i,\mathbf{c}_{sar})$
      $\ell(\phi;\sigma_i) \leftarrow \frac{1}{2}\mathbb{E}_{p_{\text{data}}(\mathbf{x})}\mathbb{E}_{\tilde{\mathbf{x}}_i\sim\mathcal{N}(\mathbf{x}_i,\sigma_i^2 I)}\left[\left\|\mathbf{s}_\phi(\tilde{\mathbf{x}}_i,\sigma_i,\mathbf{c}_{sar}) + \frac{\tilde{\mathbf{x}}_i-\mathbf{x}_i}{\sigma_i^2}\right\|_2^2\right]$
    end for
    $\mathcal{L}(\phi;\{\sigma_i\}_{i=1}^{L}) \leftarrow \frac{1}{L}\sum_{i=1}^{L}\lambda(\sigma_i)\ell(\phi;\sigma_i)$
    $\phi \leftarrow \phi - \eta\nabla_\phi\mathcal{L}(\phi;\{\sigma_i\}_{i=1}^{L})$
  end for
end for

Once the SAR-Conditioned Score Network $\mathbf{s}_{\phi^*}(\mathbf{x},\sigma,\mathbf{c}_{sar})$ is trained, the cloud-free image sampling process can be conducted by plugging the score model into the empirical PF-ODE described in Equation (4). As illustrated by Equation (8), the sampling process starts from the prior Gaussian distribution $\hat{\mathbf{x}}_T \sim \mathcal{N}(0, T^2\mathbf{I})$. An accurate estimate of $\hat{\mathbf{x}}_{t_{n-1}}^{\boldsymbol{\phi}}$ can be derived from $\mathbf{x}_{t_n}$ by executing a single discretization step with a numerical ODE solver.

$$\hat{\mathbf{x}}_{t_{n-1}}^{\boldsymbol{\phi}} := \mathbf{x}_{t_n} + (t_{n-1} - t_n)\Phi(\mathbf{x}_{t_n}, t_n, \mathbf{c}_{sar};\phi) \tag{8}$$

Here, Φ represents the numerical solver for the PF-ODE; we employ the Euler solver $\Phi(\mathbf{x},t,\mathbf{c}_{sar};\boldsymbol{\phi}) = -t\mathbf{s}_\phi(\mathbf{x},t,\mathbf{c}_{sar})$ to solve the aforementioned equation. The variable t_i denotes a specific time point, and we follow the time setting approach proposed by Karras for t_i. Specifically, the time interval $[\epsilon, T]$ is divided into $N-1$ segments, with boundaries defined by $t_1 = \epsilon < t_2 < \cdots < t_N = T$. In practice, t_i can be calculated using the following formula:

$$t_i = \left(\epsilon^{1/\rho} + \frac{i-1}{N-1}\left(T^{1/\rho} - \epsilon^{1/\rho}\right)\right)^{\rho} \tag{9}$$

For numerical stability, ϵ is set to a very small positive value, and ρ is assigned a value of 7 in this work.

The noise $\mathbf{x}_T$, sampled from the prior Gaussian distribution, along with the SAR image corresponding to the cloud-covered image, serves as input for the model. Using an ODE solver, Equation (8) is iteratively solved to produce the cloud-free image $\hat{\mathbf{x}}_0$. Analogous to image editing approaches based on diffusion models, the blending technique is employed on the cloud-covered image and the generated cloud-free image $\hat{\mathbf{x}}_0$ to derive the final de-clouded result:

$$\mathbf{x} = m \odot \hat{\mathbf{x}}_0 + (1 - m) \odot \mathbf{x}_{cloudy} \tag{10}$$

where m represents a binary mask denoting cloud-covered regions, with the value set to 1 for cloud-covered areas and 0 for cloud-free areas. $\mathbf{x}_{cloudy}$ denotes the cloud-covered image.

3.2. SAR-Conditioned Consistency Model (SCCM) via Consistency Distillation

The slow sampling speed of diffusion models represents a significant bottleneck, limiting their application in cloud removal tasks. Compared to other cloud removal methods based on DDPM models, our proposed approach, based on the score-based diffusion model, demonstrates significant improvements in sampling efficiency under the ODE framework. However, iteratively solving the ODE with a numerical solver requires multiple evaluations of the score model $\mathbf{s}_\phi(\mathbf{x}, \sigma, \mathbf{c}_{sar})$, which renders the process computationally intensive. Therefore, enhancing the speed and efficiency of cloud removal remains a crucial challenge that we aim to address.

Song et al. [42] proposed a novel addition to the family of diffusion models: consistency models, which enable single-step generation and additionally support iterative generation, balancing sample quality with computational efficiency. Inspired by the Consistency Model, we employed a consistency distillation strategy based on our pre-trained Score Model, ultimately developing a framework capable of rapid cloud removal. This approach allows for efficient inference by distilling the capabilities of the Score Model into a more streamlined process, significantly enhancing the speed of cloud removal while maintaining high-quality output. The resulting framework effectively addresses the challenges of slow sampling typically associated with diffusion models, offering a promising solution for real-time applications.

In the following section, we will introduce the consistency distillation training strategy, which builds upon the SCSBD model $\mathbf{s}_\phi(\mathbf{x}, \sigma, \mathbf{c}_{sar})$.

Figure 3 presents the consistency distillation strategy utilized in this study. The objective of the consistency distillation strategy is to establish a SAR-conditioned consistency function, denoted as $\mathcal{G}_\theta$. For the sequence of solution trajectories obtained from Equation (4), $\{\hat{\mathbf{x}}_t\}_{t\in[0,T]}$, the consistency function should satisfy the following property: for any pair $(\mathbf{x}, \mathbf{t}, \mathbf{c}_{sar})$ associated with the same PF-ODE, the consistency function should map them consistently to the identical point $\hat{x}_0$, i.e., $\mathcal{G}(\mathbf{x}_t, t, \mathbf{c}_{sar}) = \mathcal{G}(\mathbf{x}_{t'}, t', \mathbf{c}_{sar}) = \hat{x}_0$ for all $t, t' \in [0, T]$. Consequently, the distilled model allows for single-step generation, eliminating the requirement for multiple iterations typically performed by an ODE solver.

The consistency distillation training process is presented in Algorithm 2. In this process, a pair of adjacent points is sampled from the ODE trajectory, and a consistency loss is utilized to supervise their mapping back to the same origin, thereby facilitating the distillation training. Specifically, a data point $\mathbf{x} \sim p_{\text{data}}$ is initially sampled from the dataset $\mathcal{D}$, and Gaussian noise is subsequently added to it. From the SDE transition density $\mathcal{N}(\mathbf{x}, t_{n+1}^2 \boldsymbol{I})$, $\mathbf{x}_{t_{n+1}}$ is sampled. A single discretization step using an ODE numerical solver is then performed according to Equation (8) to obtain $\hat{\mathbf{x}}_{t_n}^{\phi}$, yielding a pair of adjacent points $(\hat{\mathbf{x}}_{t_n}^{\boldsymbol{\phi}}, \mathbf{x}_{t_{n+1}})$. These two points, belonging to the same ODE trajectory, should satisfy the properties of the SAR-conditioned consistency function $\mathcal{G}_\theta$. The following loss function is used to perform the distillation training of the model:

$$\mathcal{L}_{CD}^{N}(\boldsymbol{\theta}, \boldsymbol{\theta}'; \phi) = \mathbb{E}[\lambda(t_n) d(\mathcal{G}_\theta(\mathbf{x}_{t_{n+1}}, t_{n+1}, \mathbf{c}_{sar}), \mathcal{G}_{\theta'}(\hat{\mathbf{x}}_{t_n}^{\boldsymbol{\phi}}, t_n, \mathbf{c}_{sar}))] \tag{11}$$

where $n \sim \mathcal{U}[[1, N-1]]$, $\mathbf{x} \sim p_{data}$ and $\mathbf{x}_{t_{n+1}} \sim \mathcal{N}(\mathbf{x}; t_{n+1}^2 \boldsymbol{I})$. $\lambda(\cdot)$ is the positive weighting function. Furthermore, θ' refers to the running average of the parameter θ throughout the optimization process. The function $d(\cdot, \cdot)$ serves as a metric for measuring the discrepancy

in outputs between two adjacent points under the SAR-conditioned consistency function. In this paper, the mean squared error (MSE) is employed as the metric.

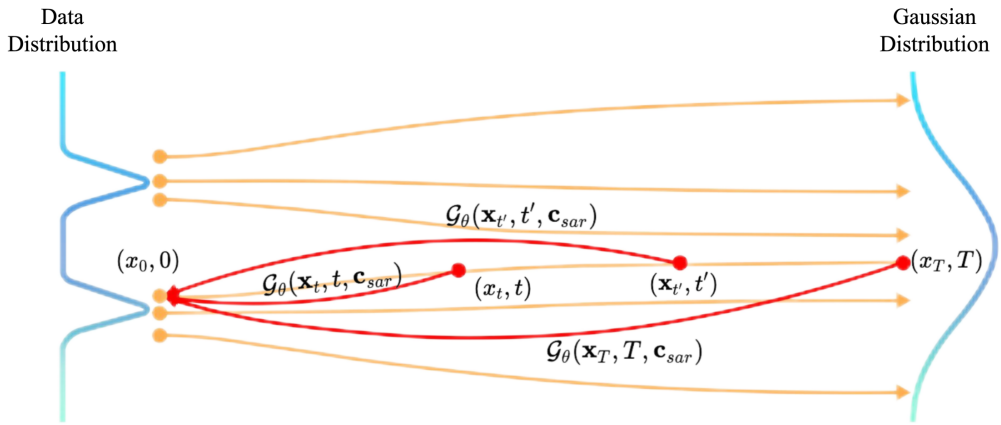


Figure 3. The SCCM is trained to map points on any trajectory of the PF ODE to the trajectory's origin.

Algorithm 2 Training SCCM via Consistency Distillation

Require: Training dataset D, initial model parameter θ, learning rate η, ODE solver $\Phi(\cdot,\cdot,\cdot;\phi)$, metric function $d(\cdot,\cdot)$, weighting function $\lambda(\cdot)$, and EMA decay rate μ
Ensure: Trained model parameter θ
1: Initialize target network parameter $\theta' \leftarrow \theta$
2: **repeat**
3: Sample $\mathbf{x} \sim D$ and $n \sim \mathcal{U}[[1, N-1]]$
4: Sample $\mathbf{x}_{t_{n+1}} \sim \mathcal{N}(\mathbf{x}; t_{n+1}^2 I)$
5: Compute $\hat{\mathbf{x}}_{t_n}^{\phi} \leftarrow \mathbf{x}_{t_{n+1}} + (t_n - t_{n+1})\Phi(\mathbf{x}_{t_{n+1}}, t_{n+1}, \mathbf{c}_{sar}; \phi)$
6: Compute loss: $\mathcal{L}(\theta, \theta'; \phi) \leftarrow \lambda(t_n) d(\mathcal{G}_\theta(\mathbf{x}_{t_{n+1}}, t_{n+1}; \mathbf{c}_{sar}), \mathcal{G}_{\theta'}(\hat{\mathbf{x}}_{t_n}^{\phi}, t_n; \mathbf{c}_{sar}))$
7: Update model parameter: $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}(\theta, \theta'; \phi)$
8: Update target network parameter: $\theta' \leftarrow \text{stop grad}(\mu\theta' + (1-\mu)\theta)$
9: **until** convergence

3.3. Progressive Denoising via Multistep Resampling for Refined Cloud Removal

Building upon SCSBD and the distillation consistency model (SCCM) described in Sections 3.1 and 3.2, this section introduces the Progressive Denoising via Multistep Resampling (PDMSR) strategy, which serves as the core inference mechanism in the CM-CR framework. PDMSR is specifically designed to enhance cloud removal fidelity while preserving high inference efficiency.

As shown in Figure 4, the PDMSR strategy begins by sampling a noise image $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$ and feeding it, along with the conditional SAR image $\mathbf{c}_{sar}$, into the consistency function $\mathcal{G}_\theta$ to generate an initial cloud-removed result $\hat{\mathbf{x}}_0$. This result is fused with the cloud-free regions of the original image $\mathbf{x}_{cloudy}$ using a binary cloud mask to obtain a hybrid image. However, directly using this one-step prediction often leads to artifacts or inconsistencies in reconstructed regions due to limited spectral guidance.

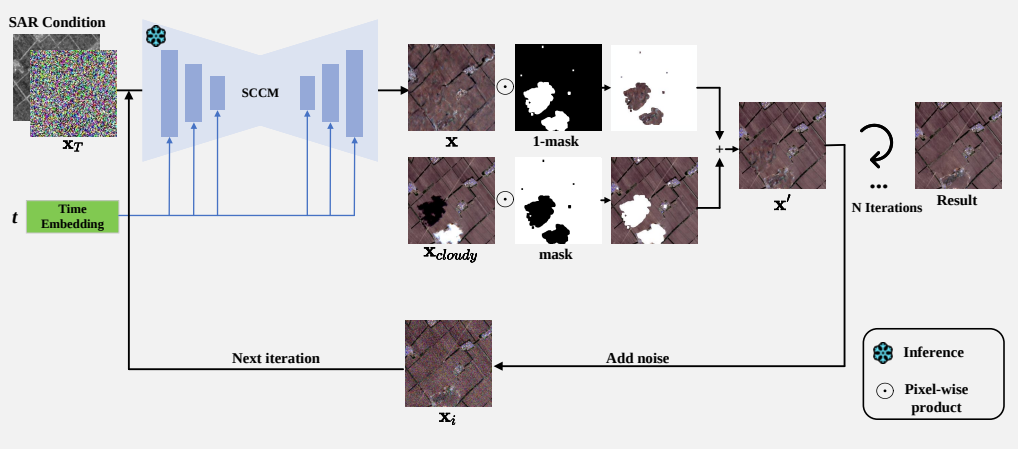


Figure 4. Illustration of the proposed PDMSR inference strategy, which progressively refines the cloud removal results by leveraging multistep denoising and re-noising guided by cloud-free regions. $\mathbf{x}_T$ denotes the initial noisy image, $\mathbf{x}_{\text{cloudy}}$ is the cloudy input, $\mathbf{x}_i$ is the intermediate noisy sample at iteration i, and $\mathbf{x}'$ is the progressively refined output.

To overcome this limitation, PDMSR introduces a multistep inference scheme that progressively reintroduces noise and refines the prediction in each step as detailed in Algorithm 3. Specifically, at each iteration, the intermediate result is perturbed with decreasing levels of Gaussian noise and then passed back through the consistency function. Moreover, cloud-free regions from the original image are preserved and fused back into the estimate after each step. This process allows the model to incorporate rich contextual and spectral priors from unoccluded areas, guiding the reconstruction of cloud-covered regions toward more semantically consistent and visually plausible outputs.

Algorithm 3 Progressive Denoising via Multistep Resampling (PDMSR)

Input: Consistency function $\mathcal{G}_\theta(\cdot,\cdot,\cdot)$, time steps $t_1 > t_2 > \cdots > t_N$, cloudy image $\mathbf{x}_{cloudy}$, binary mask m, SAR condition $\mathbf{c}_{sar}$, initial noise $\mathbf{x}_T$
$\mathbf{x} \leftarrow \mathcal{G}_{\boldsymbol{\theta}}(\mathbf{x}_T, T, \mathbf{c}_{sar})$
$\mathbf{x}' \leftarrow \mathbf{x}_{cloudy} \odot m + \mathbf{x} \odot (1 - m)$
for $i = 2$ to N **do**
  Sample $\mathbf{x}_i \sim \mathcal{N}(\mathbf{x}', (t_i^2 - \epsilon^2)\boldsymbol{I})$
  $\mathbf{x} \leftarrow \mathcal{G}_{\boldsymbol{\theta}}(\mathbf{x}_i, t_i, \mathbf{c}_{sar})$
  $\mathbf{x}' \leftarrow \mathbf{x}_{cloudy} \odot m + \mathbf{x} \odot (1 - m)$
end for
Output: $\mathbf{x}'$

Although PDMSR involves N iterative steps, each step is computationally efficient due to the single-pass property of the distilled model $\mathcal{G}_\theta$. Consequently, the overall inference speed remains significantly faster than traditional diffusion-based methods, which typically require hundreds of ODE solver steps. Our ablation studies further confirm that PDMSR substantially improves the structural consistency and visual fidelity of cloud-removed images.

4. Experimental Results

4.1. Data

The dataset utilized in this study is the publicly available SEN12MS-CR [46], comprising 169 non-overlapping regions of interest (ROIs) sampled across all inhabited continents and across various meteorological seasons. Each ROI contains a triplet of orthorectified and geo-referenced images: a cloudy Sentinel-2 optical observation, its corresponding cloud-free counterpart, and a co-registered Sentinel-1 SAR image that integrates both VV

and VH polarization channels as a dual-channel input. All three images were acquired within the same season to minimize surface variability. As illustrated in Figure 5, each ROI was further partitioned into overlapping triplet patches of 256 × 256 pixels using a stride of 128 pixels. Note that the cloudy and cloud-free Sentinel-2 images in SEN12MS-CR are not always captured on the same day. They are selected as the closest-available temporal pair within the same season to minimize surface changes, as defined by the dataset protocol. Likewise, Sentinel-1 SAR images are temporally aligned within the same seasonal window. We follow the official pairing to avoid manual alignment bias, which is a standard practice in cloud-removal studies.

The Sentinel-2 data are sourced from the Level-1C top-of-atmosphere (TOP) reflectance product, with pixel values ranging from 0 to 10,000 and including all 13 spectral bands. In contrast, Sentinel-1 images are derived from the Level-1 GRD product, acquired in Interferometric Wide (IW) swath mode with dual-polarization channels (VV and VH), and represented as backscatter coefficients in the decibel (dB) scale.

For model training, 20,000 triplet patches were selected, where only the red, green, and blue (RGB) bands from the Sentinel-2 imagery were preserved and used as input. To evaluate performance, an additional 400 representative test samples were curated according to cloud fraction and type of land cover. Specifically, 80 samples were drawn from each cloud coverage interval: 0–20%, 20–40%, 40–60%, 60–80%, and 80–100%.

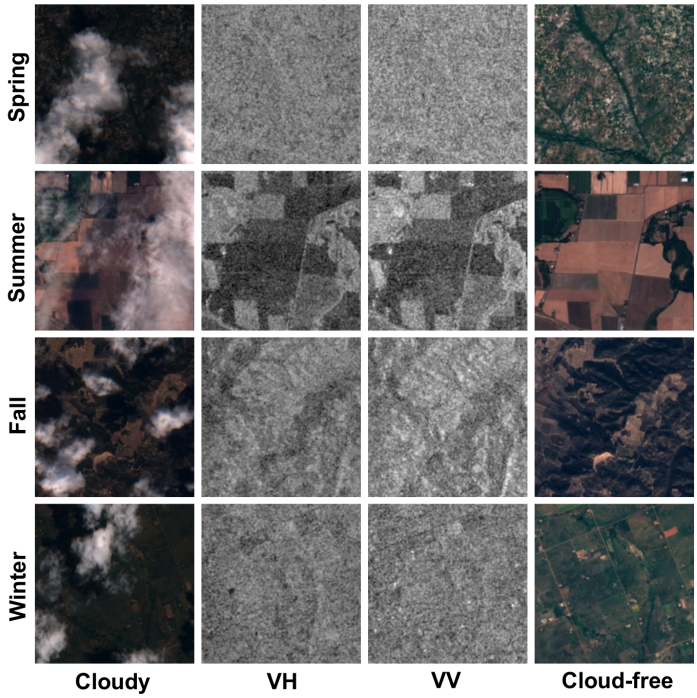


Figure 5. Representative samples from the SEN12MS-CR dataset. Rows correspond to the four meteorological seasons (spring, summer, fall, and winter), while columns represent the geo-referenced cloudy Sentinel-2 image, the Sentinel-1 SAR image with VH polarization, the SAR image with VV polarization, and a temporally matched cloud-free Sentinel-2 image.

4.2. Implementation Details

To ensure experimental reproducibility, this section presents the detailed training hyperparameter settings for SCSBD and SCCM and the evaluation metrics adopted to assess model performance.

4.2.1. Training Setting

All experiments were conducted on a high-performance computing server equipped with an AMD EPYC 7763 64-core CPU and an NVIDIA A100-SXM4-80GB GPU, running the Ubuntu operating system. The implementation was built on Python 3.10 using the PyTorch framework.

For training the SCSBD, we adopted an input image resolution of 256 × 256 pixels with a batch size of 4. The model was trained for 700,000 steps (approximately 7 days) across one A100 GPU using mixed-precision arithmetic to enhance computational efficiency. The initial learning rate was set to 0.0001, with a three-tiered exponential moving average (EMA) strategy employing decay rates of 0.999, 0.9999, and 0.9999432189950708 respectively.

During the SCCM distillation phase, we maintained identical configurations for batch size, input resolution, and computational resources while modifying specific hyperparameters: the learning rate was reduced to 0.00001, and a dropout rate of 0.1 was introduced in the teacher model. This phase utilized Heun's method as the ODE solver and incorporated LPIPS as the weighting function to quantify the differences between generated and ground-truth images.

4.2.2. Evaluation Metrics

In all experiments, we report the Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), Spectral Angle Mapper (SAM), and Mean Absolute Error (MAE) on the test set to assess the quality of the cloud removal results.

PSNR is a widely used metric in Image Quality Assessment(IQA) , where higher values generally correspond to better image quality. Its formulation is provided in Equation (12):

$$\text{PSNR} = 10 \times log_{10}\left(\frac{MAX^2}{MSE}\right) \tag{12}$$

$$MSE = \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N} \tag{13}$$

where y_i and $\hat{y}_i$ are the ground truth and predicted intensity values at the $i-$th pixel, respectively, and N is the number of pixels. MAX represents the maximum pixel intensity, set to 10,000 in this study to match the TOP reflectance scale. The Mean Squared Error (MSE), defined in Equation (13), quantifies the average squared difference between corresponding pixels.

The Structural Similarity Index Measure (SSIM) evaluates image similarity by comparing luminance, contrast, and structural information between the predicted and reference images. It is computed as:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{14}$$

where μ_x and μ_y denote the local means of images x and y, σ_x^2 and σ_y^2 are the corresponding variances, and σ_{xy} is the covariance between them. Constants C_1 and C_2 are used to stabilize the division and are derived from the dynamic range of the pixel values. The SSIM value ranges from 0 to 1, with higher values indicating greater structural similarity.

Spectral Angle Mapper (SAM) measures the spectral consistency between the predicted and reference images by calculating the angle between their spectral vectors. It is computed as:

$$\text{SAM}(x, y) = \arccos\left(\frac{\langle x, y\rangle}{\|x\|_2 \cdot \|y\|_2}\right) \tag{15}$$

where x and y denote the spectral vectors of the reference and predicted images at a given pixel location, respectively. $\langle x, y\rangle$ represents the dot product of the two vectors, and

$\|x\|_2$ and $\|y\|_2$ are their Euclidean norms. A smaller SAM value indicates higher spectral consistency, with 0 representing perfect spectral alignment.

The Mean Absolute Error (MAE) quantifies the average magnitude of pixel-wise differences between the predicted and reference images. It is insensitive to the direction of the error and directly reflects overall reconstruction accuracy. The MAE is defined as:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{16}$$

where y_i and $\hat{y}_i$ represent the ground truth and predicted values at the i-th pixel, and N is the total number of pixels. A lower MAE value indicates better performance, with 0 corresponding to a perfect reconstruction.

4.3. Cloud Removal Results

In this section, four representative terrain types—grassland, farmland, mountainous terrain, and urban areas—were selected to assess the performance of CM-CR under diverse geomorphological conditions. All data were derived from Sentinel-2 imagery within the SEN12MS-CR dataset, and true-color composites were generated using the red, green, and blue spectral bands. Despite considerable local variability in spectral and radiometric properties, as well as complex topography and significant elevation differences, the proposed CM-CR exhibited strong performance, effectively recovering most of the information obscured by cloud cover.

As shown in Figure 6, CM-CR effectively removes thick cloud cover in grassland scenes and demonstrates strong capability in reconstructing homogeneous surface areas. In particular, the red box in Figure 6a highlights that under extensive cloud coverage, the method may still introduce noticeable artifacts. Nevertheless, as illustrated in Figure 6b, even in visually complex regions, CM-CR can reliably recover uniform structures. It is worth noting that the reconstruction fidelity is influenced by the quality and structural details present in the SAR input.

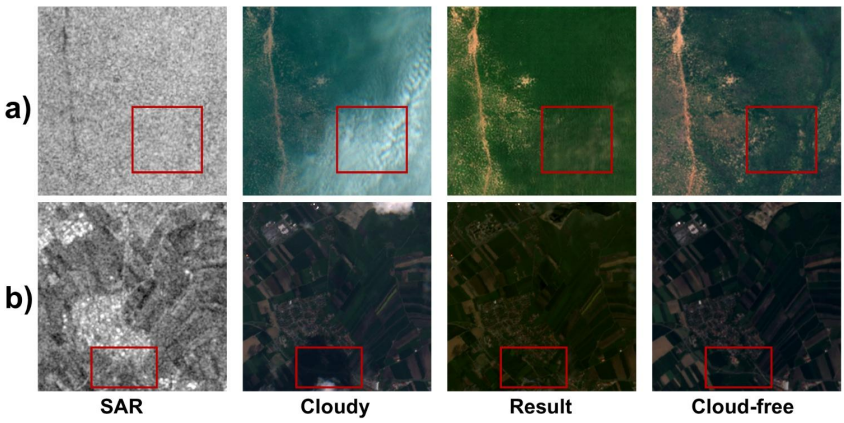


Figure 6. Cloud removal results in grassland scenes: (**a**) under extensive cloud coverage; (**b**) under complex scene conditions. Red bounding boxes indicate key regions of reconstruction.

As shown in Figure 7, CM-CR achieves notable performance in removing thick cloud cover over farmland areas. Leveraging SAR guidance, the method effectively reconstructs both structural and textural details. In the region highlighted by the red box in Figure 7a, CM-CR successfully distinguishes multiple land cover types and restores them with high fidelity. However, due to the limited accuracy of existing cloud detection algorithms, some residual cloud artifacts remain in the reconstructed output. Figure 7b further demonstrates that, even in scenes characterized by complex layouts, the model can generate content

consistent with the surrounding semantic context. Nevertheless, in regions with dense cloud coverage, the SAR input may lack sufficient structural cues, which in turn affects reconstruction accuracy and leads to noticeable deviations.

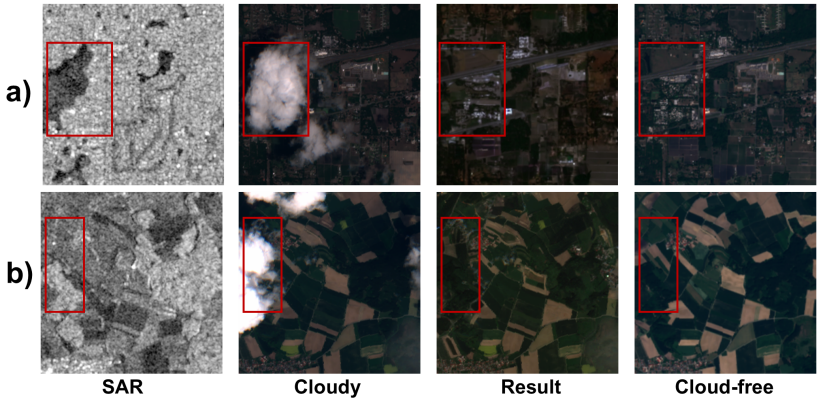


Figure 7. Cloud removal results in farmland scenes: (**a**) with multiple land cover types; (**b**) under complex layout conditions. Red bounding boxes indicate key regions of reconstruction. Red bounding boxes indicate key regions of reconstruction.

As illustrated in Figure 8, CM-CR exhibits strong reconstruction performance even in mountainous terrains with complex textures and elevation variations. In the region marked by the red box in Figure 8a, the model accurately recovers structural features present in the SAR image. However, due to the limitations of SAR in capturing fine topographic details—particularly over the central hill—noticeable discrepancies remain between the result and the reference cloud-free image. Figure 8b further illustrates the model's ability to restore mountainous structures in challenging scenarios. Nonetheless, in some areas, inconsistencies arise due to differences in acquisition time between the SAR and cloudy optical image. For example, water accumulation observed in the SAR image within valley regions is not reflected in the cloudy image, resulting in differences between the preserved cloud-free regions and reference cloud-free image.

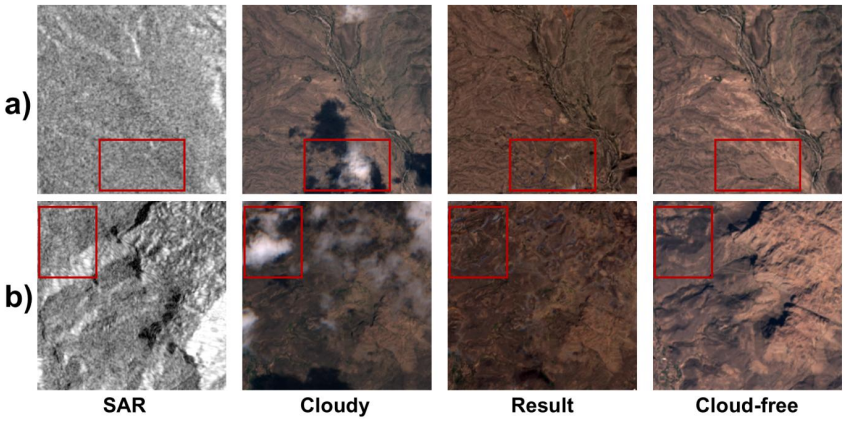


Figure 8. Cloud removal results in mountainous scenes: (**a**) with complex textures and elevation changes; (**b**) under temporal inconsistency between SAR and optical data. Red bounding boxes indicate key regions of reconstruction.

As shown in Figure 9, CM-CR demonstrates strong reconstruction performance in urban environments characterized by complex layouts and high-frequency textures. In Figure 9a, the model attempts to recover dense road networks and their local branching structures from a heavily cloud-obscured input. While it successfully reconstructs portions that are clearly reflected in the SAR image, a downward-branching segment is missing

due to its poor visibility and ambiguity in the SAR signal. In contrast, Figure 9b illustrates a more favorable outcome. Despite thick cloud cover in the optical observation, CM-CR effectively restores primary roadways and building layouts with notable consistency. The clearer structural cues in the SAR image significantly enhance reconstruction quality in occluded regions. Nevertheless, minor discrepancies persist in areas with highly textured content, underscoring the inherent challenges of urban reconstruction under severely degraded observations.

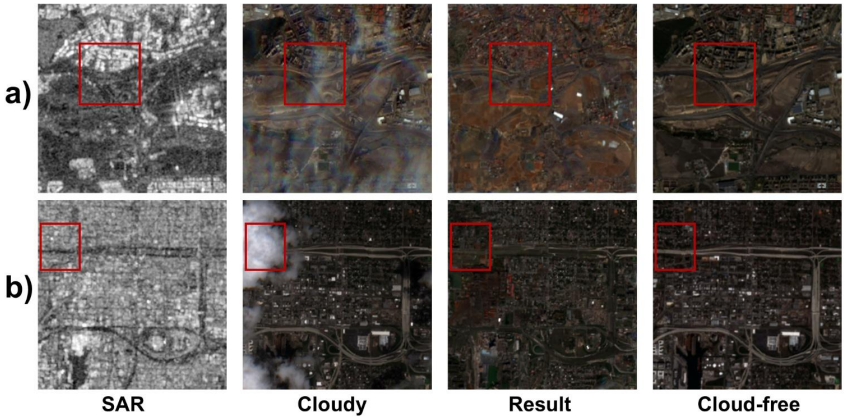


Figure 9. Cloud removal results in urban scenes: (**a**) with dense road networks; (**b**) under heavy cloud occlusion with clear structural guidance. Red bounding boxes indicate key regions of reconstruction.

4.4. Comparisons with State-of-the-Art Methods

We conducted a comparative analysis of the proposed model against existing methods—including the discriminative baselines Dsen2-CR [46], GLF-CR [47], and UnCR [48], as well as diffusion-based approaches DiffCR [36] and Repaint [49]—evaluating both cloud removal performance and computational efficiency. All models were trained on the same set of 20,000 triplet samples and tested on the same evaluation dataset. For DiffCR and Repaint, the number of sampling steps was kept consistent with their original implementations, both using 200 steps. Since our CM-CR model achieves the reconstruction results shown in Table 1 with only 13 PDMSR steps, we further included a qualitative comparison where Repaint's sampling steps were artificially reduced to 13—denoted as Repaint-13—to enable a more direct comparison with our approach. Since the GLF-CR method uses 128 × 128 inputs while all other methods use 256 × 256 images, we combined four non-overlapping 128 × 128 patches to form a single 256 × 256 image, ensuring fairness when calculating reconstruction quality metrics. Additionally, the "Time" reported for GLF-CR reflects the total time needed to process these four 128 × 128 patches.

Table 1. Quantitative comparison of cloud removal methods. Bold indicates the best, and underlined indicates the second-best results.

Method	PSNR (dB) ↑	SSIM ↑	SAM ↓	MAE (TOP) ↓	Params (M)	Time (s/img) ↓
Dsen2-CR [46]	32.214	0.906	0.048	244.03	75.9	**0.062**
GLF-CR [47]	<u>33.159</u>	<u>0.922</u>	<u>0.043</u>	<u>221.08</u>	73.1	0.095
UnCR [48]	30.686	0.897	0.062	290.53	97.1	<u>0.088</u>
DiffCR [36]	29.317	0.861	0.105	364.85	87.5	2.218
Repaint [49]	32.508	0.907	0.054	246.05	34.62	83.062
CM-CR (Ours)	**34.191**	**0.935**	**0.042**	**202.69**	2730.71	2.023

We first conducted quantitative comparative experiments with current state-of-the-art cloud removal methods. Table 1 summarizes the detailed image reconstruction quality metrics, including the average PSNR, SSIM, SAM, and MAE. In addition, it reports model performance metrics, specifically the number of parameters (Params) and the time required

to reconstruct a single image (Time). The experimental results show that our method consistently outperforms all other approaches in evaluation metrics, fully demonstrating its effectiveness for the cloud removal task. Specifically, compared to the second-best method, our approach achieves notable improvements: the PSNR increases by 3.1%, the SSIM improves by 1.4%, while the SAM and MAE decrease by 2.3% and 8.3%, respectively. Although the inference time of CM-CR is still slightly longer than that of discriminative models, its parameter count is approximately 28 times greater than that of UnCR, the largest among the discriminative baselines. This larger parameter scale provides CM-CR with stronger feature extraction capacity, enabling it to learn more robust representations and achieve superior reconstruction performance. Notably, compared to the Repaint method, which adopts a standard DDPM sampling strategy, CM-CR achieves an impressive 40× reduction in per-image processing time while using about 80 times more parameters. Even when compared with the efficient diffusion-based model DiffCR, CM-CR still improves per-image processing speed by approximately 8.8% despite having about 30 times more parameters. These results demonstrate the high sampling efficiency of CM-CR.

We next conducted qualitative comparative experiments with current state-of-the-art cloud removal methods. Figures 10–13 present qualitative comparison results of different methods across various scenes. As shown in Figure 10, the comparison illustrates how each method, assisted by SAR, performs in reconstructing scene structures. Figure 10a depicts a scene with simple structural layouts, whereas Figure 10b shows a more complex structural scenario. The discriminative models Dsen2-CR, GLF-CR, and UnCR can partially recover missing structures; however, due to their limited capacity to restore fine details, the reconstructed regions appear overly smoothed. DiffCR fails to preserve structural consistency, leading to incomplete or distorted reconstructions. Repaint-13, which reduces the sampling steps from 200 to 13, results in insufficient denoising and noticeable residual noise. In contrast, both Repaint and CM-CR successfully reconstruct clear and well-defined structural elements, with the restored regions exhibiting abundant fine-grained details.

Figure 11 presents a scene containing both textural and structural features, where Figure 11a depicts a simple texture–structure scenario and Figure 11b illustrates a more complex one. Taking the red-marked region in Figure 11a as an example, in the simple setting, the discriminative models can accurately distinguish edges and texture structures, producing recognizable shapes with coherent and naturally smooth textures. However, the reconstruction by UnCR shows noticeable blurring in the restored areas. In addition, both DiffCR and Repaint-13 fail to preserve structural consistency. While Repaint and CM-CR can both recover accurate structural information, Repaint exhibits color deviations in the restored textures. By contrast, the reconstruction by CM-CR yields smoother textures with better structural fidelity. In Figure 11b, the red-marked area highlights a region with complex textures and structural content. In this more challenging scenario, Dsen2-CR, GLF-CR, UnCR, DiffCR, and Repaint-13 are unable to faithfully reconstruct either the structures or the detailed textures. Although Repaint partially recovers complex structures and textures, its accuracy remains limited. In comparison, our method, CM-CR, can more precisely reconstruct both complex structural elements and multi-scale texture details.

Figure 12 shows reconstruction results for scenes containing complex high-frequency details. Specifically, Figure 12a presents a scene with a small amount of such information, while Figure 12b illustrates a scene rich in complex high-frequency features. As shown in the red-marked area of Figure 12a, in a scene with limited complex high-frequency information, the Dsen2-CR, GLF-CR, and UnCR methods can only recover coarse structural outlines and fail to capture fine details. DiffCR and Repaint-13 fail to produce meaningful reconstructions altogether. The Repaint method achieves relatively better results but still loses typical high-frequency elements such as buildings. In contrast, our method not

only restores the structural content but also fully preserves high-frequency features like buildings. This advantage arises because the standard DDPM sampling mechanism used by Repaint tends to generate content that aligns with surrounding semantic context, making it prone to missing localized high-frequency details. By comparison, our CM-CR method requires only 13 sampling steps, significantly reducing the risk of neglecting such high-frequency information. In the red-marked area of Figure 12b, the cloud-covered region contains dense buildings and other complex high-frequency details. In this case, models such as Dsen2-CR, GLF-CR, UnCR, DiffCR, and Repaint-13 all fail to effectively reconstruct such regions rich in fine-grained structures. Among them, the Repaint model can only partially recover structural information, whereas our CM-CR model clearly reconstructs the dense and complex high-frequency content. This result demonstrates the superior generative capability and reconstruction effectiveness of CM-CR in challenging scenarios.

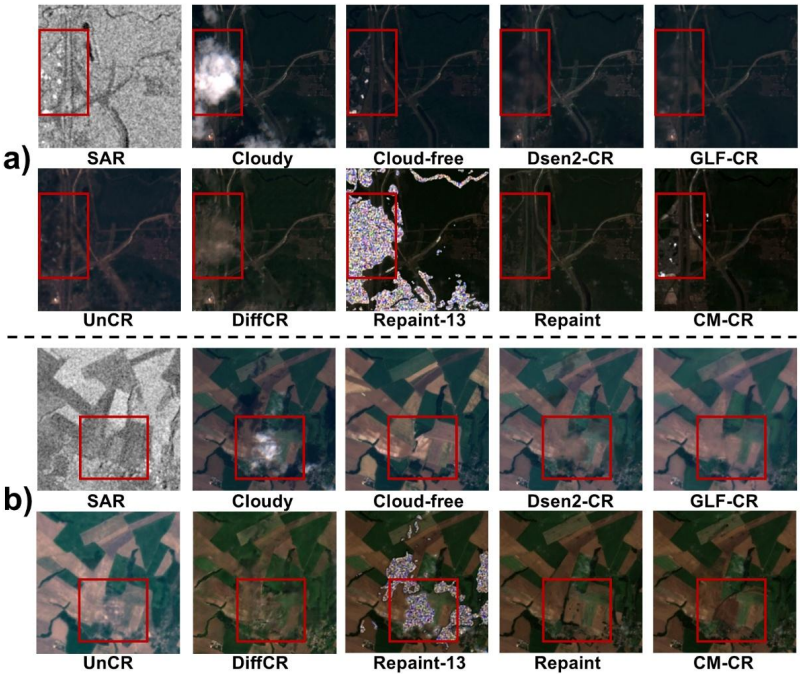


Figure 10. Qualitative comparison of scene structural reconstruction capabilities for cloud removal algorithms. (**a**) shows a test scene with simple geometric structures, while (**b**) depicts a scene with more complex topological patterns. Red-marked areas highlight key regions of reconstruction. For better visual inspection, zooming in is recommended to examine differences in structural reconstruction quality.

Figure 13 provides a qualitative comparison of various cloud removal algorithms applied to scenes with prominent high-frequency details and mixed-frequency characteristics. As shown in the red-marked area of Figure 13a, the scene features dominant high-frequency details, specifically bright building rooftops. Here, Dsen2-CR, GLF-CR, and UnCR recover only coarse structural outlines and miss fine rooftop details. DiffCR and Repaint-13 fail to preserve these features, leading to incomplete reconstructions with visible artifacts. While Repaint achieves better structural recovery, it still shows color inconsistencies and blurred rooftop edges. In contrast, our CM-CR model clearly restores the bright rooftops with sharp structures and fine-grained details, demonstrating strong localized high-frequency reconstruction. In Figure 13b, the scene combines low-frequency textures with high-frequency details. For this mixed region, Dsen2-CR, GLF-CR, UnCR, DiffCR, and Repaint-13 struggle to maintain texture continuity and structural sharpness, often producing blurred or inconsistent areas. Repaint partially recovers textures and boundaries but retains artifacts. By comparison, CM-CR generates coherent results that

preserve both textures and embedded high-frequency details, showing its effectiveness in complex mixed-frequency scenarios.

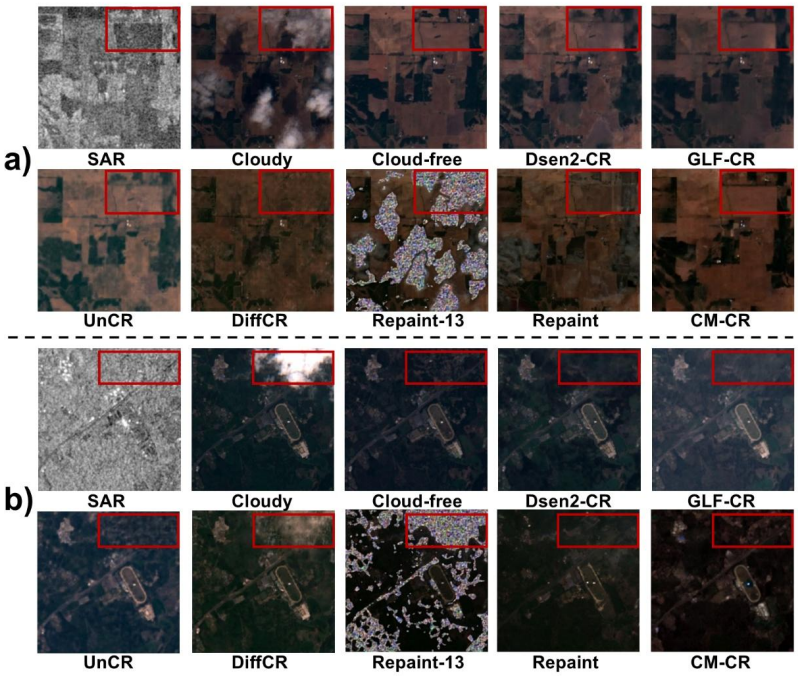


Figure 11. Qualitative comparison of the structural and textural co-reconstruction performance of cloud removal algorithms. (**a**) shows a test scene with simple geometric structures and uniform texture features, while (**b**) depicts a representative scene containing complex topological structures and multi-scale textural details. Red-marked areas highlight key regions of reconstruction. For better visual inspection, zooming in is recommended to examine differences in structural reconstruction quality.

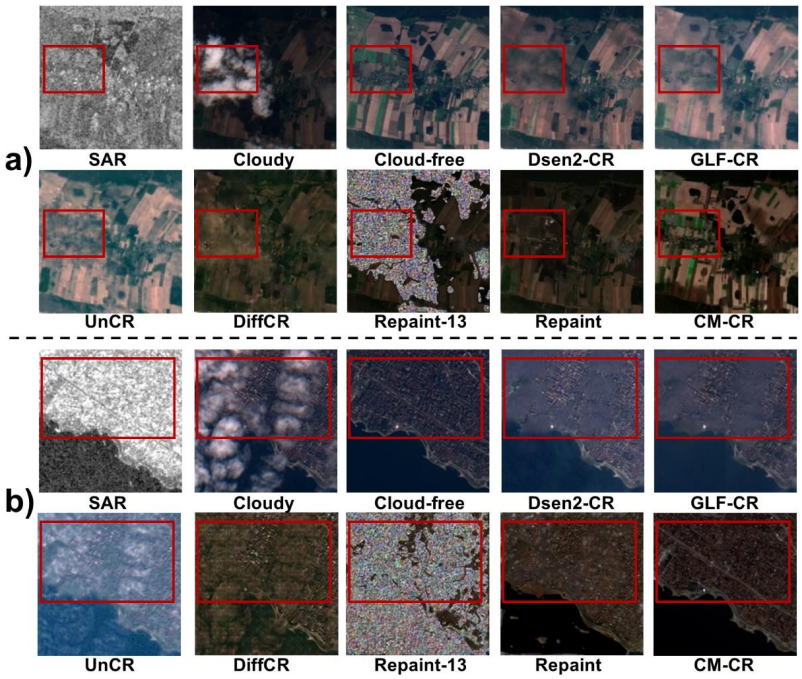


Figure 12. Qualitative comparison of the reconstruction performance of cloud removal algorithms in scenes with varying levels of high-frequency information. (**a**) shows a relatively simple scene with sparse high-frequency details, while (**b**) depicts a more complex scene rich in dense high-frequency content. Red-marked areas highlight key regions of reconstruction. For better visual inspection, zooming in is recommended to examine differences in structural reconstruction quality.

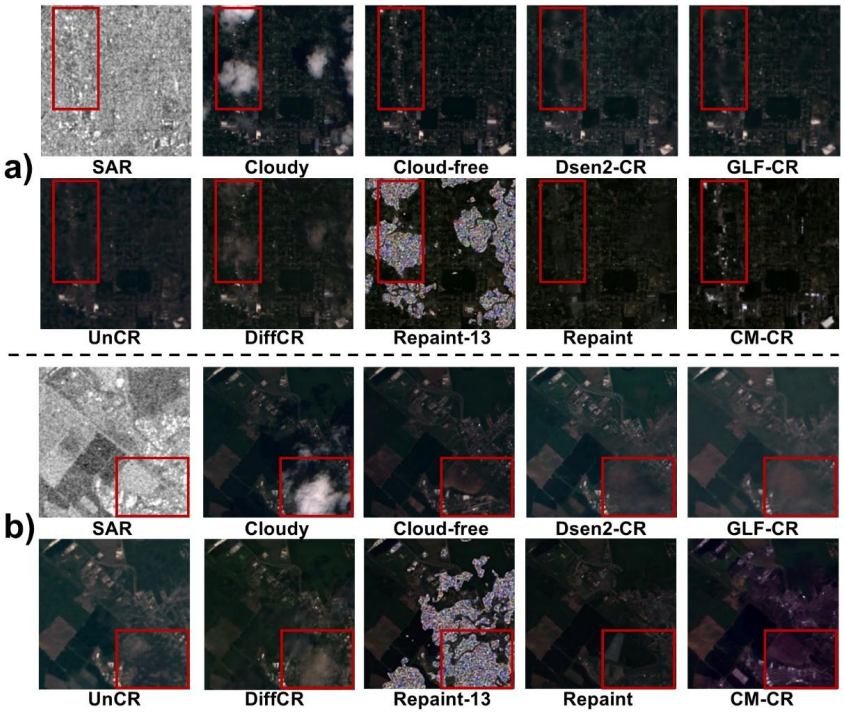


Figure 13. Qualitative comparison of the reconstruction performance of cloud removal algorithms in scenes featuring dominant high-frequency details and mixed-frequency characteristics. (**a**) shows a test scene where bright building rooftops serve as distinct high-frequency components. (**b**) presents a composite scene where low-frequency textures and high-frequency details coexist. Red-marked areas highlight key regions of reconstruction. For better visual inspection, zooming in is recommended to examine differences in structural reconstruction quality.

In summary, the qualitative results across diverse scenes confirm that all baseline methods struggle to simultaneously restore structural, textural, and spectral details under dense cloud cover. Discriminative models often oversmooth fine structures, while DiffCR and Repaint-13 fail to fully recover high-frequency elements due to limited sampling steps. The standard Repaint method with longer sampling partially improves structural quality but still shows color inconsistencies and high computational cost. Compared to these baselines, CM-CR consistently achieves clearer structures, richer textures, and better local detail preservation with a much faster sampling process, demonstrating its strong potential for robust cloud removal in complex scenarios.

4.5. Ablation Study

4.5.1. The Effectiveness of SAR Image

To verify the effectiveness of SAR condition in cloud removal, both quantitative and qualitative comparisons are conducted, as shown in Table 2 and Figure 14, respectively.

Quantitatively, the inclusion of SAR auxiliary information significantly enhances reconstruction quality across all evaluation metrics. Specifically, compared to the w/o SAR setting, the $w/$SAR configuration improves the PSNR from 28.431 to 34.191 and the SSIM from 0.794 to 0.9346. Meanwhile, the SAM and MAE values drop from 0.177 to 0.042 and from 379.04 to 202.69, respectively. These results clearly demonstrate that SAR conditioning improves not only the fidelity but also the structural and spectral consistency of the reconstructed images.

Qualitatively, as illustrated in Figure 14, reconstructions without SAR exhibit pronounced randomness in the generated textures and structures, especially in the heavily cloud-covered regions, leading to unrealistic or distorted outcomes. In contrast, with SAR assistance, the model effectively reconstructs terrain-aligned structures and fine-grained

textures, guided by backscattering priors inherent in SAR data. This improvement can be attributed to the auxiliary structural and backscattering information provided by SAR imagery, which effectively compensates for the information lost due to cloud occlusion in the optical domain.

These results jointly confirm the substantial benefit of incorporating SAR data into the cloud removal framework, enabling more accurate, consistent, and reliable image reconstruction under challenging cloud conditions.

Table 2. Ablation study on the impact of SAR conditioning for cloud removal.

Method	PSNR (↑)	SSIM (↑)	SAM (↓)	MAE (↓)
w/*o* SAR	28.431	0.794	0.177	379.04
w/SAR	34.191	0.9346	0.042	202.69

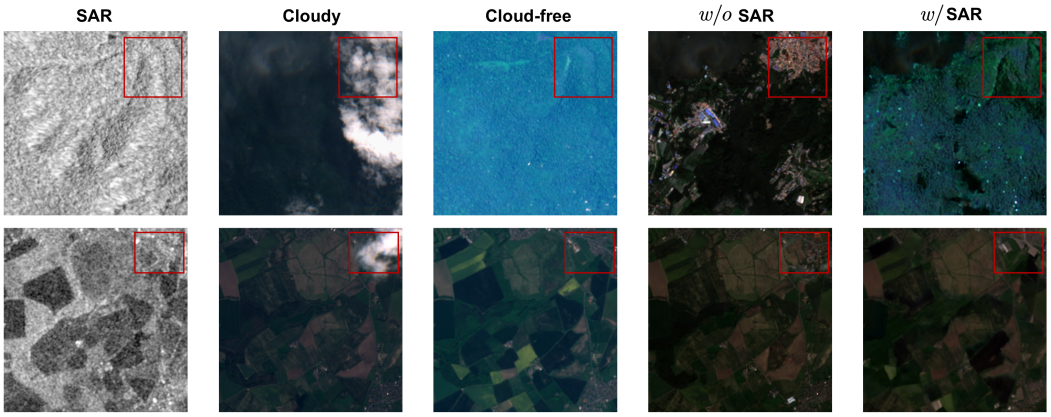


Figure 14. Qualitative comparison results illustrating the impact of SAR are shown. The label "*w*/*o* SAR" denotes cloud removal outcomes without incorporating SAR imagery, whereas "*w*/SAR" indicates results generated with SAR auxiliary.

4.5.2. Numbers of PDMSR

To investigate the impact of the number of resampling steps in the PDMSR strategy, we conduct a visual analysis of the reconstruction quality under different sampling iterations. The results are presented in Figure 15, which shows the cloud removal progression from the initial noise (index 1) and the one-step coarse prediction (index 2), followed by PDMSR refinement results corresponding to 1 to 18 PDMSR steps (indices 3–20).

It is observed that single-step sampling (index 2) struggles to produce satisfactory reconstruction results, often leading to incomplete or noisy reconstructions in heavily cloud-covered regions. As the number of PDMSR steps increases, the reconstructed regions gradually improve in visual quality, exhibiting enhanced structural continuity and reduced artifacts. Notably, by the 13th PDMSR step (index 15), CM-CR is already able to produce high-quality reconstructions that closely resemble the underlying terrain. Further increasing the number of steps beyond this point yields diminishing returns, with negligible improvements observed in visual fidelity or detail enhancement. As summarized in Table 3, reconstruction accuracy improves gradually with more PDMSR steps and stabilizes around 13 steps, while runtime grows nearly linearly with the number of steps. This supports our choice of roughly 13 PDMSR steps as a practical efficiency–quality trade-off.

These findings suggest that while increasing the number of PDMSR steps enhances reconstruction effectiveness, there exists an optimal threshold (around Step 13) beyond which the benefits plateau. This insight provides a practical guideline for balancing inference efficiency and reconstruction quality in real-world applications of the CM-CR method.

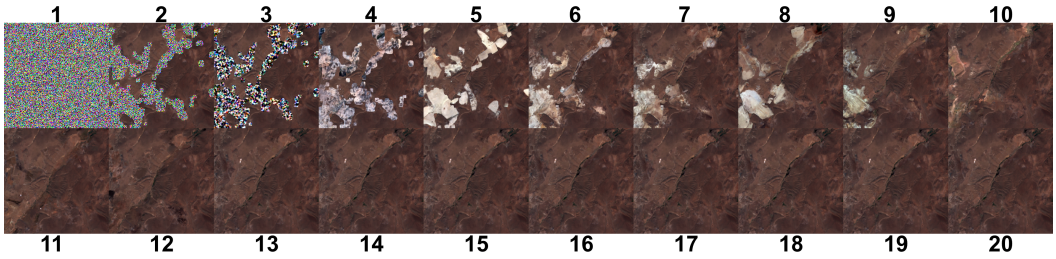


Figure 15. Cloud removal progression with PDMSR steps. Index 1 denotes the initial noise; index 2 denotes the coarse prediction (one-step student inference); indices 3–20 denote results after applying PDMSR with 1 to 18 refinement steps.

Table 3. Effect of PDMSR steps on reconstruction quality and runtime. Note that "Step = 1" corresponds to the first PDMSR refinement output in Figure 15 (i.e., index 3 in that figure), following the initial noise (index 1) and the one-step coarse prediction (index 2).

Steps	PSNR ↑	SSIM ↑	SAM ↓	MAE (TOP) ↓	Time (s/img) ↓
1	29.125	0.858	0.1080	372.0	0.311
5	31.642	0.905	0.0550	245.0	0.933
9	33.172	0.928	0.0465	210.0	1.555
13	34.191	0.935	0.0420	202.69	2.178
17	34.213	0.936	0.0418	201.8	2.801

5. Discussion

In this section, we focus on two key questions: (1) How does CM-CR perform under low-quality SAR input conditions? (2) How does its cloud removal performance compare to baseline models across different levels of cloud coverage? Figure 16 illustrates two representative failure cases of reconstruction due to poor SAR quality. Figure 17 provides a quantitative comparison of image quality trends across methods under cloud coverage ratios ranging from 0% to 100%.

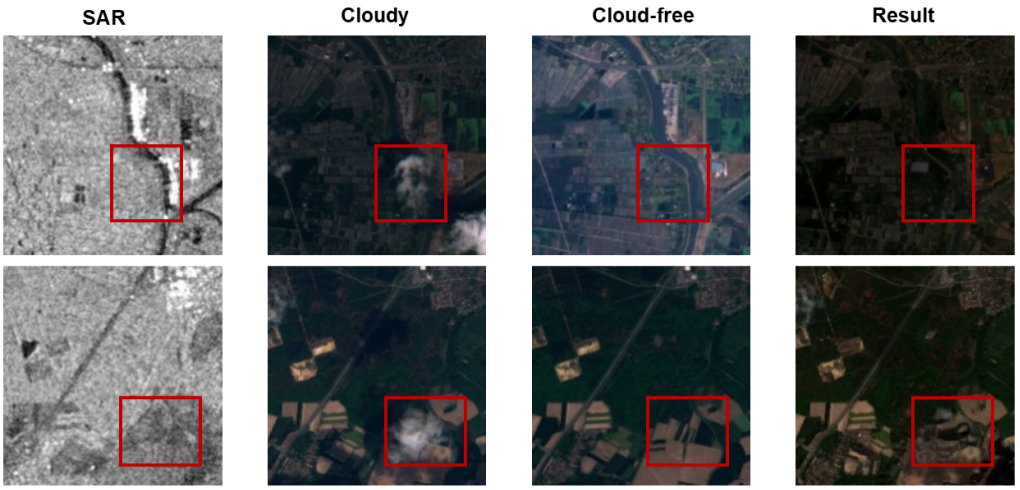


Figure 16. Qualitative performance analysis under low-quality SAR image conditions.

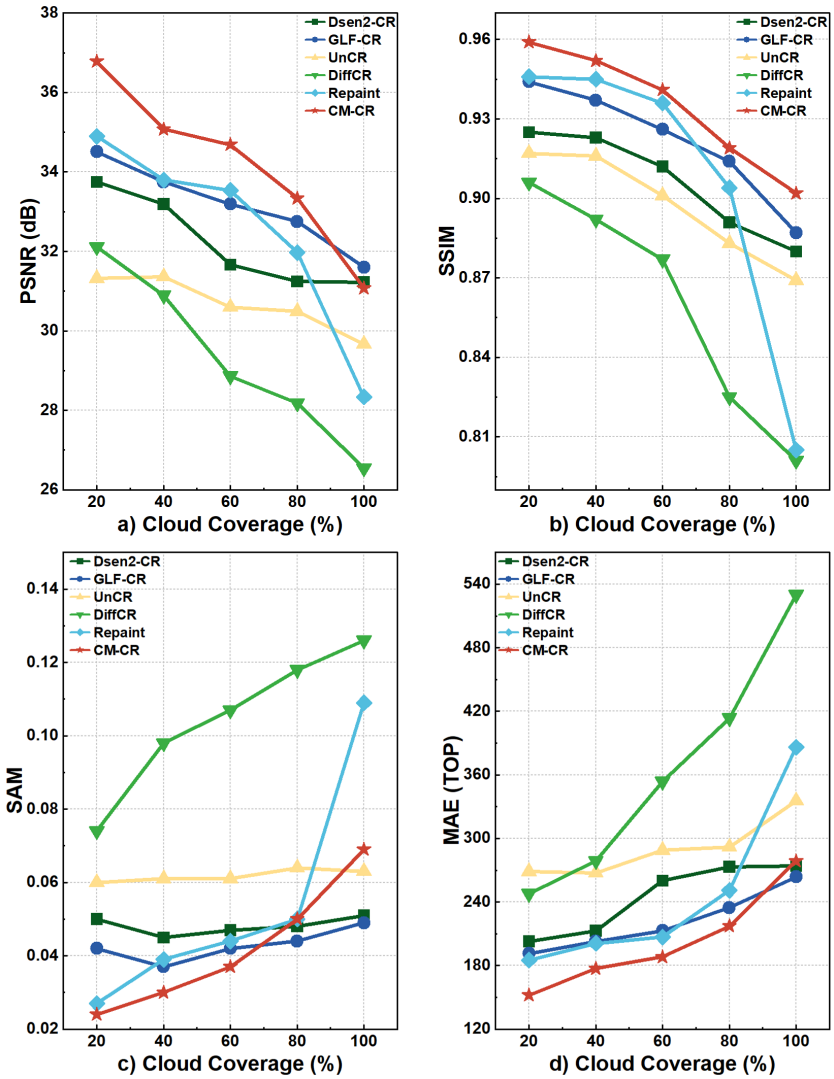


Figure 17. Image quality metrics for different methods vary with cloud coverage levels. Subfigures (**a**–**d**) show the PSNR, SSIM, SAM, and MAE variations across different cloud coverage ranges, respectively.

5.1. Performance Analysis Under Low-Quality SAR Image

The reconstruction process of CM-CR is highly dependent on the auxiliary information provided by SAR imagery, which places high demands on the quality of the SAR input. As discussed in Section 4.4, we found that when SAR images accurately provide both low-frequency and high-frequency information, CM-CR consistently achieves favorable reconstruction results. However, in real-world applications, the quality of SAR imagery often cannot be maintained at a consistently high level and may fail to deliver reliable low- and high-frequency information. For example, as shown in the red-marked area of the first row in Figure 16, the SAR image appears as a homogeneous region, whereas the cloud-free reference image from another time indicates that the same region contains complex structural details. This suggests that the SAR image has lost structural information in that area, resulting in CM-CR failing to recover the corresponding structural features. This demonstrates the limitation of CM-CR in its strong dependence on SAR image quality.

As shown in the second row of Figure 16, the reconstruction results produced by CM-CR exhibit some randomness; nevertheless, some details are still recovered, and the visual quality of the reconstructed region remains acceptable. This may be attributed to the fact that although the efficient sampling strategy adopted by CM-CR enhances the model's

generalization capability to some extent, it is not universally applicable to all scenarios, particularly when the auxiliary information provided by SAR imagery is relatively limited.

In the SEN12MS-CR dataset, the SAR images have already undergone standard preprocessing, including radiometric calibration, speckle suppression, and terrain correction. Therefore, the "low-quality" SAR samples observed in this study mainly originate from intrinsic acquisition limitations rather than post-processing imperfections. Typical cases include geometric distortions (layover or shadowing) and weak backscatter from smooth or specular surfaces, where structural details are genuinely absent from the SAR signal and cannot be recovered by conventional enhancement. To address this limitation, incorporating complementary information from multi-temporal optical observations or auxiliary modalities such as Landsat-8, which provide reliable spectral reflectance and temporal consistency, could effectively compensate for the missing structural or textural cues. From the modeling perspective, designing modules that better capture the contextual and semantic information of optical imagery—such as cross-modal attention or spatial–semantic reasoning blocks—may further enhance reconstruction performance when SAR guidance is degraded. These strategies indicate a promising direction toward a unified, multi-source framework for robust cloud removal under challenging SAR conditions.

5.2. Analysis of Different Cloud Coverage

Figure 17 illustrates the quantitative trends of reconstruction quality for different cloud removal algorithms as cloud coverage varies.

Figure 17a shows that as cloud coverage increases, the PSNR of all algorithms steadily decreases, mainly because greater cloud coverage raises the complexity of recovering surface information. Among the discriminative models, GLF-CR performs better than Dsen2-CR and UnCR under light to moderate cloud coverage (coverage < 60%), consistently maintaining higher PSNR values. However, once the coverage exceeds this threshold, its performance deteriorates more rapidly. DiffCR consistently shows relatively low reconstruction quality across all cloud coverage levels, with PSNR values noticeably lower than those of other baselines. Repaint and CM-CR, which rely on explicit cloud region masks during reconstruction, are more affected under high-cloud-coverage conditions (coverage > 60%). Notably, when cloud coverage approaches 100%, the PSNR of CM-CR even falls below that of GLF-CR and Dsen2-CR. This indicates that explicit mask-based models may become less robust in extremely dense cloud scenarios where accurate mask estimation and guidance become more challenging.

In Figure 17b, a similar decreasing trend is observed for the SSIM. As cloud coverage increases, the structural similarity achieved by the discriminative methods (Dsen2-CR, GLF-CR, and UnCR) gradually declines. DiffCR and Repaint also exhibit a pronounced drop in SSIM under heavy-cloud conditions, indicating limitations in maintaining structural consistency. In contrast, the CM-CR method demonstrates a stronger capability to preserve structural information, with its SSIM values consistently leading the other methods. This result highlights that CM-CR can better leverage the auxiliary information provided by SAR imagery to maintain structural coherence even under dense cloud coverage.

Figure 17c shows the evaluation results for the SAM. For this metric, a lower SAM value indicates better spectral consistency. The discriminative models Dsen2-CR, GLF-CR, and UnCR demonstrate relatively stable SAM values, with only slight increases as cloud coverage increases. In contrast, the SAM value for the Repaint model rises significantly when cloud coverage exceeds 80%, reflecting poorer spectral consistency under dense clouds. However, although the CM-CR model exhibits clear advantages when cloud coverage is below 80%, its SAM value surpasses that of the UnCR model when cloud coverage

exceeds this level. This indicates that our method may face limitations in maintaining spectral consistency under very high-cloud-coverage conditions.

Figure 17d shows the variation in MAE with increasing cloud coverage. Among the discriminative methods, UnCR generally produces higher MAE values under thick-cloud conditions. As cloud coverage increases, DiffCR shows a pronounced increase in error. While the Repaint method maintains moderate error levels, it does not match the lower MAE achieved by CM-CR. The CM-CR method consistently yields the lowest MAE values, with only slight increases over GLF-CR and Dsen2-CR when cloud coverage exceeds 80%. This highlights the advantage of CM-CR in minimizing per-pixel reconstruction errors across different cloud scenarios.

Overall, the results under different cloud coverage levels show that all methods degrade as cloud coverage increases, reflecting the increasing challenge of recovering structural and spectral details under dense clouds. Among the discriminative models, methods like GLF-CR and Dsen2-CR remain relatively stable at low to moderate coverage but decline more sharply beyond 60%. DiffCR consistently performs worse than the discriminative baselines across all metrics, while Repaint achieves better reconstruction quality than DiffCR and most discriminative models under moderate coverage but suffers under very high cloud amounts due to its dependence on accurate masks. CM-CR maintains clear advantages across most metrics and coverage levels, though its performance gap narrows when coverage exceeds 80%, highlighting that further improvements are needed to handle near-complete occlusion.

6. Conclusions

In this study, we proposed CM-CR, a novel consistency model-based architecture that effectively removes thick clouds from optical remote sensing imagery. By integrating SAR information and adopting a consistency distillation strategy, the model achieves high-quality reconstruction while significantly improving sampling efficiency compared to standard diffusion models. Extensive experiments conducted on the SEN12MS-CR benchmark demonstrate that CM-CR consistently outperforms mainstream discriminative and diffusion-based cloud removal methods in terms of PSNR, SSIM, SAM, and MAE metrics. The key contribution of this work lies in dividing the complex reconstruction process into a rapid coarse prediction and a progressive refinement stage, which balances reconstruction fidelity and inference speed. In the future, further research will focus on enhancing the model's generalization capability for different satellite data sources, exploring multi-modal data fusion, and reducing reliance on explicit cloud masks. This work provides a promising direction for improving the usability of optical remote sensing data under challenging cloud conditions.

Author Contributions: Conceptualization, Q.H.; methodology, Q.H., B.H. and Y.L.; software, Q.H. and B.H.; validation, Q.H. and B.H.; formal analysis, Q.H. and B.H.; investigation, Q.H.; resources, Q.H. and Y.L.; data curation, Q.H. and B.H.; writing—original draft, Q.H., B.H. and Y.L.; writing—review and editing, Q.H., B.H. and Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (62271400).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The SEN12MS-CR dataset is available online at https://patricktum.github.io/cloud_removal/sen12mscr/(accessed on 10 November 2025).

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

References

1. He, J.; Wang, W.; Fu, M.; Wang, Y. Insights into global visibility patterns: Spatiotemporal distributions revealed by satellite remote sensing. *J. Clean. Prod.* **2024**, *468*, 143069. [CrossRef]
2. Victor, N.; Maddikunta, P.K.R.; Mary, D.R.K.; Murugan, R.; Chengoden, R.; Gadekallu, T.R.; Rakesh, N.; Zhu, Y.; Paek, J. Remote Sensing for Agriculture in the Era of Industry 5.0–A survey. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 5920–5945. [CrossRef]
3. Jia, P.; Chen, C.; Zhang, D.; Sang, Y.; Zhang, L. Semantic segmentation of deep learning remote sensing images based on band combination principle: Application in urban planning and land use. *Comput. Commun.* **2024**, *217*, 97–106. [CrossRef]
4. Ahmad, R. Smart remote sensing network for disaster management: an overview. *Telecommun. Syst.* **2024**, *87*, 213–237. [CrossRef]
5. Gao, J.; Yuan, Q.; Li, J.; Zhang, H.; Su, X. Cloud Removal with Fusion of High Resolution Optical and SAR Images Using Generative Adversarial Networks. *Remote Sens.* **2020**, *12*, 191. [CrossRef]
6. Xu, M.; Jia, X.; Pickering, M.; Jia, S. Thin cloud removal from optical remote sensing images using the noise-adjusted principal components transform. *ISPRS J. Photogramm. Remote Sens.* **2019**, *149*, 215–225. [CrossRef]
7. Shen, H.; Li, X.; Cheng, Q.; Zeng, C.; Yang, G.; Li, H.; Zhang, L. Missing information reconstruction of remote sensing data: A technical review. *IEEE Geosci. Remote Sens. Mag.* **2015**, *3*, 61–85. [CrossRef]
8. Rossi, R.E.; Dungan, J.L.; Beck, L.R. Kriging in the shadows: geostatistical interpolation for remote sensing. *Remote Sens. Environ.* **1994**, *49*, 32–40. [CrossRef]
9. Shen, H.; Zhang, L. A MAP-based algorithm for destriping and inpainting of remotely sensed images. *IEEE Trans. Geosci. Remote Sens.* **2008**, *47*, 1492–1502. [CrossRef]
10. Maalouf, A.; Carré, P.; Augereau, B.; Fernandez-Maloigne, C. A bandelet-based inpainting technique for clouds removal from remotely sensed images. *IEEE Trans. Geosci. Remote Sens.* **2009**, *47*, 2363–2371. [CrossRef]
11. Lorenzi, L.; Melgani, F.; Mercier, G. Missing-area reconstruction in multispectral images under a compressive sensing perspective. *IEEE Trans. Geosci. Remote Sens.* **2013**, *51*, 3998–4008. [CrossRef]
12. Wang, L.; Wang, Q.; Tong, X.; Atkinson, P.M. Mst-net: A general deep learning model for thick cloud removal from optical images. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5612518. [CrossRef]
13. Sekrecka, A.; Karwowska, K. Classical vs. Machine Learning-Based Inpainting for Enhanced Classification of Remote Sensing Image. *Remote Sens.* **2025**, *17*, 1305. [CrossRef]
14. Chen, Y.; Chen, M.; He, W.; Zeng, J.; Huang, M.; Zheng, Y.B. Thick cloud removal in multitemporal remote sensing images via low-rank regularized self-supervised network. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5506613. [CrossRef]
15. Zhou, H.; Wang, Y.; Liu, W.; Tao, D.; Ma, W.; Liu, B. MSC-GAN: A Multi-Stream Complementary Generative Adversarial Network with Grouping Learning for Multitemporal Cloud Removal. *IEEE Trans. Geosci. Remote Sens.* **2024**, *63*, 5400117. [CrossRef]
16. Zou, X.; Li, K.; Xing, J.; Tao, P.; Cui, Y. PMAA: A progressive multi-scale attention autoencoder model for high-performance cloud removal from multi-temporal satellite imagery. *arXiv* **2023**, arXiv:2303.16565
17. Zhang, Q.; Yuan, Q.; Li, Z.; Sun, F.; Zhang, L. Combined deep prior with low-rank tensor SVD for thick cloud removal in multitemporal images. *ISPRS J. Photogramm. Remote Sens.* **2021**, *177*, 161–173. [CrossRef]
18. Peng, H.; Huang, T.Z.; Zhao, X.L.; Lin, J.; Wu, W.H.; Li, L.Y. Deep Domain Fidelity and Low-Rank Tensor Ring Regularization for Thick Cloud Removal of Multitemporal Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5409314. [CrossRef]
19. Li, J.; Wang, Y.; Sheng, Q.; Wu, Z.; Wang, B.; Ling, X.; Liu, X.; Du, Y.; Gao, F.; Camps-Valls, G.; et al. CloudRuler: Rule-based transformer for cloud removal in Landsat images. *Remote Sens. Environ.* **2025**, *328*, 114913. [CrossRef]
20. Wang, L.; Wang, Q.; Atkinson, P.M. Thick cloud removal of Landsat time-series using convolutional LSTM with embedded residual modules. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5404816. [CrossRef]
21. Xia, Y.; He, W.; Huang, Q.; Yin, G.; Liu, W.; Zhang, H. CRformer: Multi-modal data fusion to reconstruct cloud-free optical imagery. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *128*, 103793. [CrossRef]
22. Duan, C.; Belgiu, M.; Stein, A. Efficient Cloud Removal Network for Satellite Images Using SAR-optical Image Fusion. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 6008605. [CrossRef]
23. Ebel, P.; Meraner, A.; Schmitt, M.; Zhu, X.X. Multisensor data fusion for cloud removal in global and all-season sentinel-2 imagery. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 5866–5878. [CrossRef]
24. Bermudez, J.D.; Happ, P.N.; Oliveira, D.A.B.; Feitosa, R.Q. SAR to optical image synthesis for cloud removal with generative adversarial networks. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2018**, *4*, 5–11. [CrossRef]
25. Bhambani, K.; Takalikar, M. DeCloud GAN: An Advanced Generative Adversarial Network for Removing Cloud Cover in Optical Remote Sensing Imagery. In Proceedings of the 2021 4th International Conference on Computational Intelligence and Intelligent Systems, Tokyo, Japan, 20–22 November 2021; pp. 25–30. [CrossRef]
26. Zhu, H.; Wang, Z.; Han, L.; Xu, M.; Li, W.; Liu, Q.; Liu, S.; Du, B. TSMCF: Transformer-Based SAR and Multi-Spectral Cross-Attention Fusion for Cloud Removal. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 6710–6720. [CrossRef]

27. Zhang, W.; Mei, J.; Wang, Y. Dmdiff: A dual-branch multimodal conditional guided diffusion model for cloud removal through sar-optical data fusion. *Remote Sens.* **2025**, *17*, 965. [CrossRef]
28. Singh, P.; Komodakis, N. Cloud-gan: Cloud removal for sentinel-2 imagery using a cyclic consistent generative adversarial networks. In Proceedings of the IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018; pp. 1772–1775. [CrossRef]
29. Zhao, Y.; Shen, S.; Hu, J.; Li, Y.; Pan, J. Cloud removal using multimodal GAN with adversarial consistency loss. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 8015605. [CrossRef]
30. Darbaghshahi, F.N.; Mohammadi, M.R.; Soryani, M. Cloud removal in remote sensing images using generative adversarial networks and SAR-to-optical image translation. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 4105309. [CrossRef]
31. Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6840–6851.
32. Dhariwal, P.; Nichol, A. Diffusion models beat gans on image synthesis. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 8780–8794.
33. Yang, S.; Chen, X.; Liao, J. Uni-paint: A unified framework for multimodal image inpainting with pretrained diffusion model. In Proceedings of the 31st ACM International Conference on Multimedia, Ottawa, ON, Canada, 29 October–3 November 2023; pp. 3190–3199. [CrossRef]
34. Li, H.; Yang, Y.; Chang, M.; Chen, S.; Feng, H.; Xu, Z.; Li, Q.; Chen, Y. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing* **2022**, *479*, 47–59. [CrossRef]
35. Chung, J.; Hyun, S.; Heo, J.P. Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 8795–8805. [CrossRef]
36. Zou, X.; Li, K.; Xing, J.; Zhang, Y.; Wang, S.; Jin, L.; Tao, P. DiffCR: A Fast Conditional Diffusion Framework for Cloud Removal From Optical Satellite Images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5612014. [CrossRef]
37. Zhao, X.; Jia, K. Cloud removal in remote sensing using sequential-based diffusion models. *Remote Sens.* **2023**, *15*, 2861. [CrossRef]
38. Jing, R.; Duan, F.; Lu, F.; Zhang, M.; Zhao, W. Denoising diffusion probabilistic feature-based network for cloud removal in Sentinel-2 imagery. *Remote Sens.* **2023**, *15*, 2217. [CrossRef]
39. Wang, M.; Song, Y.; Wei, P.; Xian, X.; Shi, Y.; Lin, L. IDF-CR: Iterative Diffusion Process for Divide-and-Conquer Cloud Removal in Remote-sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5615014. [CrossRef]
40. Sui, J.; Ma, Y.; Yang, W.; Zhang, X.; Pun, M.O.; Liu, J. Diffusion Enhancement for Cloud Removal in Ultra-Resolution Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5615014. [CrossRef]
41. Liu, Y.; Li, W.; Guan, J.; Zhou, S.; Zhang, Y. Effective cloud removal for remote sensing images by an improved mean-reverting denoising model with elucidated design space. In Proceedings of the Computer Vision and Pattern Recognition Conference, Nashville, TN, USA, 11–15 June 2025; pp. 17851–17861. [CrossRef]
42. Song, Y.; Dhariwal, P.; Chen, M.; Sutskever, I. Consistency Models. In Proceedings of the 40th International Conference on Machine Learning. PMLR, Honolulu, HI, USA, 23–29 July 2023; Volume 202; pp. 32211–32252.
43. Song, Y.; Sohl-Dickstein, J.; Kingma, D.P.; Kumar, A.; Ermon, S.; Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv* **2020**, arXiv:2011.13456.
44. Karras, T.; Aittala, M.; Aila, T.; Laine, S. Elucidating the design space of diffusion-based generative models. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 26565–26577.
45. 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, Munich, Germany, 5–9 October 2015; Springer: Cham, Switzerland, 2015; pp. 234–241. [CrossRef]
46. Meraner, A.; Ebel, P.; Zhu, X.X.; Schmitt, M. Cloud removal in Sentinel-2 imagery using a deep residual neural network and SAR-optical data fusion. *ISPRS J. Photogramm. Remote Sens.* **2020**, *166*, 333–346. [CrossRef]
47. Xu, F.; Shi, Y.; Ebel, P.; Yu, L.; Xia, G.S.; Yang, W.; Zhu, X.X. GLF-CR: SAR-enhanced cloud removal with global–local fusion. *ISPRS J. Photogramm. Remote Sens.* **2022**, *192*, 268–278. [CrossRef]
48. Ebel, P.; Garnot, V.S.F.; Schmitt, M.; Wegner, J.D.; Zhu, X.X. UnCRtainTS: Uncertainty quantification for cloud removal in optical satellite time series. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 2086–2096. [CrossRef]
49. Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; Van Gool, L. Repaint: Inpainting using denoising diffusion probabilistic models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 11461–11471. [CrossRef]

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.