remote sensing

MDPI

Article

AFR-CR: An Adaptive Frequency Domain Feature Reconstruction-Based Method for Cloud Removal via SAR-Assisted Remote Sensing Image Fusion

Xiufang Zhou [1,2,3], Qirui Fang [1,3,4], Xunqiang Gong [1,2,3,*], Shuting Yang [1,3], Tieding Lu [1,3], Yuting Wan [2], Ailong Ma [2] and Yanfei Zhong [2]

1 National Key Laboratory of Uranium Resources Exploration-Mining and Nuclear Remote Sensing, East China University of Technology, Nanchang 330013, China; zhouxf@ecut.edu.cn (X.Z.); 2022110383@ecut.edu.cn (Q.F.); styang02@ecut.edu.cn (S.Y.); tdlu@ecut.edu.cn (T.L.)
2 State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China; wanyuting@whu.edu.cn (Y.W.); maailong007@whu.edu.cn (A.M.); zhongyanfei@whu.edu.cn (Y.Z.)
3 Key Laboratory of Mine Environmental Monitoring and Improving Around Poyang Lake of Ministry of Natural Resources, East China University of Technology, Nanchang 330013, China
4 The Nanjing Krystal Vision Technologies, Ltd., Nanjing 210000, China
* Correspondence: xqgong1988@ecut.edu.cn

Highlights

What are the main findings?

- We propose a novel SAR-optical fusion method for cloud removal, featuring an adaptive frequency-domain framework that effectively decouples and cross-reconstructs multimodal features.
- The proposed method achieves top performance on the M3R-CR dataset, outperforming eight competing methods across all key metrics.

What are the implications of the main findings?

- The method effectively resolves the issue of blurred edges in reconstructed images, significantly enhancing the usability of optical data under cloudy conditions.
- It provides a more reliable source of cloud-free imagery for critical applications like disaster monitoring, crop classification, and land cover mapping.

Abstract

Optical imagery is often contaminated by clouds to varying degrees, which greatly affects the interpretation and analysis of images. Synthetic Aperture Radar (SAR) possesses the characteristic of penetrating clouds and mist, and a common strategy in SAR-assisted cloud removal involves fusing SAR and optical data and leveraging deep learning networks to reconstruct cloud-free optical imagery. However, these methods do not fully consider the characteristics of the frequency domain when processing feature integration, resulting in blurred edges of the generated cloudless optical images. Therefore, an adaptive frequency domain feature reconstruction-based cloud removal method is proposed to solve the problem. The proposed method comprises four key sequential stages. First, shallow features are extracted by fusing optical and SAR images. Second, a Transformer-based encoder captures multi-scale semantic features. Subsequently, the Frequency Domain Decoupling Module (FDDM) is employed. Utilizing a Dynamic Mask Generation mechanism, it explicitly decomposes features into low-frequency structures and high-frequency details, effectively suppressing cloud interference while preserving surface textures. Finally, robust information interaction is facilitated by the Cross-Frequency Reconstruction Module (CFRM)

Check for updates

Academic Editors: Shouhang Du, Maofan Zhao, Jancsó Tamás, Jiahao Wu and Abhishek Singh

Received: 3 December 2025
Revised: 23 December 2025
Accepted: 25 December 2025
Published: 8 January 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

via transposed cross-attention, ensuring precise fusion and reconstruction. Experimental evaluation on the M3R-CR dataset confirms that the proposed approach achieves the best results on all four evaluated metrics, surpassing the performance of the eight other State-of-the-Art methods. It has demonstrated its effectiveness and advanced capabilities in the task of SAR-optical fusion for cloud removal.

Keywords: optical remote sensing images; synthetic aperture radar; cloud removal; adaptive frequency domain feature reconstruction

1. Introduction

Remote sensing technology assumes a pivotal role in diverse domains, including environmental monitoring, resource management, and climate research [1–6]. However, optical remote sensing images are often disturbed by clouds, which limits the application of remote sensing images. Therefore, the ability to remove clouds is an indispensable step for enhancing the integrity and application value of optical remote sensing data [7,8]. Cloud coverage obscures surface details in remote sensing imagery, thereby introducing substantial uncertainty and complexity into the cloud removal task [9]. Synthetic Aperture Radar (SAR) [10], performing observations unhindered by weather or cloud cover [11], can capture backscattered signals from the surface of the Earth, thus alleviating the challenges in the cloud removal process to some extent [12]. By conducting fusion analysis of SAR data and optical remote sensing images, multi-source data support can be provided for the reconstruction of surface information in cloud coverage regions, thereby further improving the accuracy and reliability of cloud removal results [13,14].

Cloud removal is the process of reconstructing obscured regions in optical satellite imagery caused by cloud cover [15–17]. Early methods mainly relied on the assumption that the area occluded by clouds has the same statistical and geometric structure as other unoccluded regions in the image. For instance, cloud removal technology based on Bandelet transform and multi-scale geometric grouping is used to analyze the geometric flow of the unoccluded area in the image, such as the direction and continuity of edges and textures [18,19]. The geometric structural information is propagated to cloud-occluded regions, thus enabling the reconstruction of missing data. By minimizing the energy function, this technique interpolates pixel data from the cloud-surrounding regions to the occluded regions along the geometric flow direction, achieving the objective of cloud removal. This kind of method regards cloud removal as an interpolation task and uses the pixel information around the cloud to predict the data occluded by the cloud. However, in scenarios where the cloud-covered area is excessively extensive, it remains challenging to restore high-quality cloud-free imagery solely by leveraging the information contained within a single image [20,21].

To address this problem, several studies have proposed multi-temporal data-based methods, which utilize complementary imagery acquired at temporally adjacent or phenologically similar dates to recover the missing information in the current image [22,23]. Among these methods, the cloud removal method based on spatiotemporal fusion incorporates auxiliary cloud-free images, captures the temporal variation characteristics of land cover via an improved spatiotemporal data fusion model, and integrates a residual correction strategy to solve the restoration challenge encountered in scenarios characterized by large-scale cloud interference and significant land cover changes [24,25]. To address the challenge of reconstructing remote sensing images obscured by dense cloud cover, Zhang et al. [23] stacked time-series images into third-order tensors, exploited the low-rank

characteristics of data, introduced a 3D convolutional neural network to construct deep spatiotemporal priors, and integrated low-rank tensors decomposition with deep learning. This integrated approach thereby achieves efficient restoration of regions with large-scale cloud coverage. However, under prolonged cloudy conditions, the acquisition of pristine reference imagery from temporally proximal dates poses significant challenges. Therefore, using SAR data as a complementary information source to reconstruct cloud-affected optical imagery has become a research hotspot [26,27].

To achieve a more robust cloud removal effect, the SAR-OPT method performs targeted adjustments and in-depth development on the pix2pix architecture to build a novel cloud removal network [28]. The underlying principle of this method is to combine SAR and multispectral images to generate cloudless optical images, which shows excellent performance in detail restoration and global consistency maintenance. Compared with models that only rely on a single optical data, the reconstruction quality of this method is greatly improved. Faced with the challenge that cloud occlusion affects the quality of optical images, the DSen2-CR method takes the stitched results of Sentinel-1 and cloud-corrupted Sentinel-2 data as input, and adopts a depth residual network as its fundamental network architecture to predict the target cloud-free optical images [29]. Meanwhile, an adaptive cloud loss function is introduced to better preserve the original image information. The GLF-CR method, proposed by Xu et al., addresses the requirement of cloud removal that necessitates balancing global and local consistency, and incorporates both global and local image information throughout the reconstruction process [30]. This method adopts SAR image data as the guidance for optical image reconstruction, while simultaneously utilizing optics to guide SAR denoising; this bidirectional guidance mechanism ultimately enhances the cloud removal performance. To solve the problem of collaborative cloud removal of low-resolution and high-resolution data, the Align-CR approach leverages low-resolution SAR imagery as a guide to reconstruct high-resolution optical information, and effectively complements the misalignment between multi-modal, multi-resolution sources through the implicit feature map alignment inherent in its reconstruction process [31]. More recently, HF-T2CR [26] further advanced this direction by proposing a unified framework that deeply leverages SAR guidance to enhance restoration fidelity. Despite the benefits of multi-modal fusion, existing SAR-based methods are predominantly constructed using convolutional architectures. Consequently, they often struggle to capture long-range dependencies and global context due to their inherently limited receptive fields.

In recent years, Transformer-based architectures have achieved remarkable progress in various image tasks due to their strong capability in modeling these long-range dependencies. Advanced general restoration backbones, such as ESSAformer [32] and RKformer [33], utilize efficient attention mechanisms to capture global features. ESSAformer introduces a spectral-spatial attention mechanism for hyperspectral image super-resolution, while RKformer integrates the Runge–Kutta method for small target detection. Although these methods demonstrate powerful feature extraction capabilities, their direct application to cloud removal is often suboptimal. They primarily focus on spatial or general spectral feature extraction and overlook the specific degradation characteristics of clouds in the frequency domain. Specifically, clouds primarily act as low-frequency additive noise, while the degradation of ground details manifests as a loss of high-frequency structures.

To address the limitations of spatial-only models and generic Transformers, researchers have recently begun to explore the potential of the frequency domain. CR-Famba [34] incorporates frequency information to enhance restoration quality by modeling spectral degradation. However, despite its improvements, CR-Famba relies on fixed spectral bases and lacks effective cross-modal guidance. It typically treats frequency components globally and fails to adaptively decouple frequency features based on varying cloud thickness.

Consequently, existing methods have yet to fully exploit the synergy between SAR-derived structural priors and adaptive cross-frequency interactions, which is essential for restoring clear textures in severely occluded regions.

To address the above problems, a remote sensing image cloud removal method combining the spatial domain and the frequency domain is proposed in this paper. The main contributions of this study are as follows:

1. Spatial-Frequency Interactive Framework: We propose a novel AFR-CR architecture that explicitly models the interaction between spatial and frequency domains. Unlike conventional spatial-only methods, this framework integrates a multi-scale Transformer encoder with a frequency-aware reconstruction mechanism, utilizing SAR priors to effectively disentangle cloud artifacts from surface details for robust recovery.
2. The Adaptive Frequency Domain Decoupling Module (FDDM): In this study, the designed FDDM utilizes a Dynamic Mask Generation (DMG) mechanism to spectrally decompose cloud-containing images. The module adaptively extracts low-frequency global information and high-frequency detailed information at different frequency levels, ensuring precise feature separation even under uneven cloud cover.
3. The Cross-Frequency Domain Feature Reconstruction Module (CFRM): In this paper, a CFRM is proposed, which is integrated with a transposed cross-attention mechanism. The module optimizes the feature mapping process through effective cross-frequency domain information interaction, facilitating robust texture restoration in complex scenarios.

Extensive experiments on the M3R-CR dataset demonstrate that the proposed AFR-CR achieves State-of-the-Art performance, outperforming eight advanced methods in terms of Peak Signal-to-Noise Ratio (PSNR) [35], Structure Similarity Index Measure (SSIM) [36], Mean Absolute Error (MAE) [37], and Spectral Angle Mapper (SAM) [38].

This paper is organized as follows: Section 2 presents the detailed model design of the AFR-CR method. Section 3 presents the experimental results, including experimental design, experimental analysis, ablation experiments, and computational efficiency analysis. Section 4 discusses the analysis at different cloud cover levels. Finally, Section 5 provides the conclusions.

2. Methods

The proposed method mainly includes the multi-modal image feature fusion and the adaptive frequency feature reconstruction. In the stage of multi-modal image feature fusion, a multi-channel feature fusion method for multi-layer semantic and texture features is employed, which is based on the mosaicking of optical and SAR images and combined with the encoder–decoder structure. In the multi-modal image feature fusion stage, a multi-channel feature fusion method for multi-layer semantic and texture features is employed, which is based on the mosaicking of optical and SAR images and combined with the encoder–decoder structure. In the stage of adaptive frequency feature reconstruction, a stepwise reconstruction strategy based on multi-level encoding and decoding is put forward. Aiming at solving the problem that different degrees of cloud amount will affect the image content with different frequency bands, this study specially designs an Adaptive Frequency Reconstruction Block (AFRB). This block decomposes the input features into their constituent low- and high-frequency parts, subsequently applying targeted modulation to guide the reconstruction of the respective spectral sub-bands. AFRB includes two key components, the FDDM and the CFRM.

The experimental procedure primarily comprises four components: multi-channel feature extraction, multi-layer feature encoding, adaptive frequency feature reconstruction,

and high-quality image output. The proposed AFR-CR framework and its processing flow are shown in Figure 1. The implementation steps are as follows.

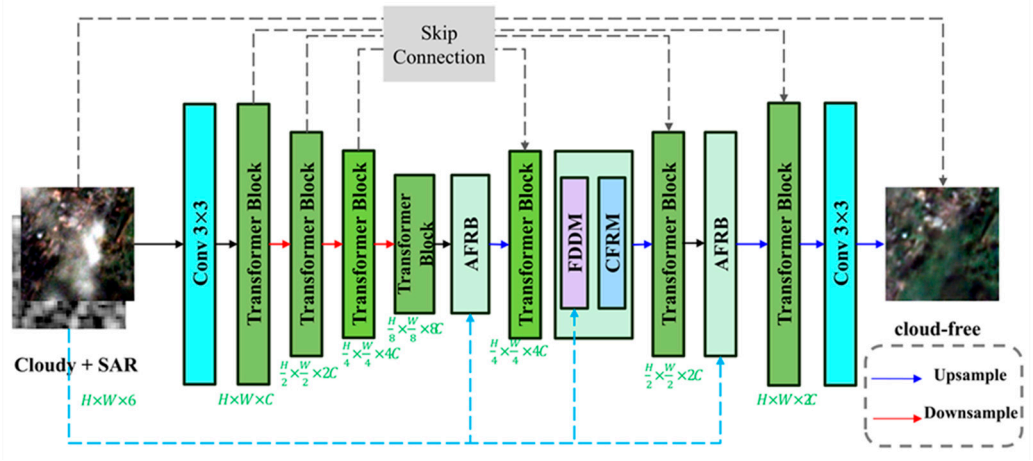

Figure 1. AFR-CR frame diagram. Dotted lines represent skip connections.

1. Multi-Channel Feature Extraction: The cloud-containing optical images $I_{\text{opt}} \in \mathbb{R}^{H\times W\times 4}$ are spliced with their corresponding SAR images $I_{\text{SAR}} \in \mathbb{R}^{H\times W\times 2}$ to form the multichannel features $I \in \mathbb{R}^{H\times W\times 6}$; and then, shallow features $Y_0 \in \mathbb{R}^{H\times W\times C}$ are extracted via the 3×3 convolutional layer.
2. Multi-Layer Feature Encoding: The four-layer Transformer Block (TB) of the encoder is employed to refine the initially fused features. Concurrently, high-resolution features Y_0 are continuously extracted downward to derive low-resolution latent representations $Y_l \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times 8C}$ thereby enabling the extraction of multi-layer semantic and texture features across diverse resolutions and scales.
3. Adaptive Frequency Feature Reconstruction: Firstly, a gradual reconstruction process is implemented, spanning from the top layer of the encoder to the bottom layer of the decoder, with the integration of the interleaved Transformer Block and AFRB. Subsequently, the FDDM segregates low-frequency and high-frequency information, followed by the execution of feature enhancement operations. Finally, the CFRM leverages convolution kernels of varying scales to process the aforementioned low-frequency and high-frequency information, guaranteeing a balance between feature details and the overall contour. The above operations not only improve the model's feature reconstruction ability but also realize the feature interaction between the frequency domain and the spatial domain.
4. High-Quality Image Output: The features output by the bottom layer of the decoder are integrated into high-quality cloud-free reconstructed imagery through the additional 3×3 convolutional layers, achieving a deep integration of the optical and SAR data, which effectively supplements the information of cloud occlusion regions and removes speckle noise.

2.1. Frequency Domain Decoupling Module (FDDM)

The FDDM is presented in Figure 2a. This module is mainly responsible for two operations: frequency domain conversion and dynamic mask generation in the first step, and frequency feature decoupling and multi-convolution cross-attention fusion in the second step.

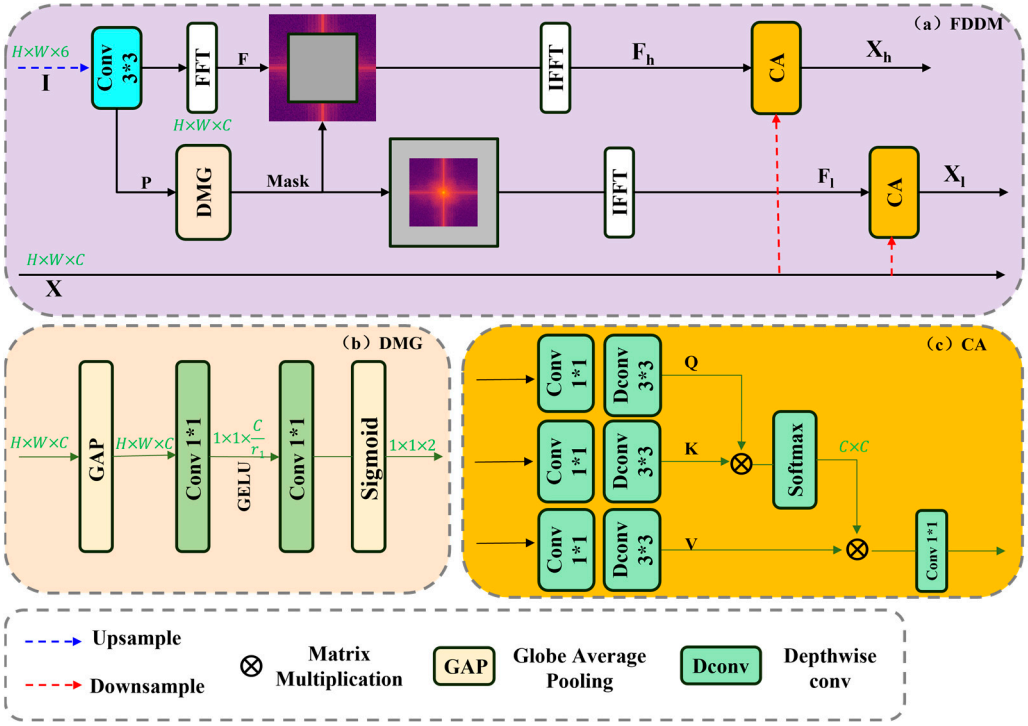

Figure 2. Frequency domain decoupling module.

2.1.1. Frequency Domain Conversion and Dynamic Mask Generation

First, the input optical images and SAR images are mosaicked to obtain the multi-channel feature I, and both I and the intermediate feature X are fed as inputs. FDDM extracts different frequency features from I to cope with diverse cloud cover scenarios. In the step of frequency domain conversion, the FDDM processes the degraded images through 3×3 convolutional layers to augment its channel dimensionality to expand the channel capacity, making it match the channel count of the input features. Subsequently, the FDDM transforms these output features into frequency-domain representations via the Fast Fourier Transform (FFT). Because there may be uneven cloud coverage in different areas of the imagery, FDDM needs to adaptively decompose the input features into their different frequency components. Therefore, a lightweight Dynamic Mask Generation (DMG) module is designed in this study to generate adaptive two-dimensional masks, which are used to separate the spectrum information of input images. The mask can be adaptively adjusted in accordance with the cloud coverage within the image, effectively improving the cloud removal effect.

The DMG is shown in Figure 2b. The projected features are first mapped into vectors via the global average pooling operator. Secondly, two adaptive factors ranging from 0 to 1 are generated via two 1×1 convolutional layers equipped with GELU activation functions; these factors define the mask size by multiplying the spectral width and spectral height. This dynamically generated mask can not only effectively distinguish low-frequency and high-frequency features in optical images and SAR images, but also generate different masks according to different regions covered by clouds, ensuring that each area achieves effective spectrum separation and cloud removal processing. The generation process of masks can be expressed as:

$$[\alpha, \beta] = \delta\left(W_2^{1\times1}\left(\sigma\left(W_1^{1\times1}(GAP_s(\mathbf{P}))\right)\right)\right) \quad (1)$$

In the formula, α and β are dynamically generated adaptive factors that define the size of the dynamic mask. $\mathbf{P}$ is the input image feature; GAP_s represents the spatial global

average pooling; σ denotes the GELU activation function; δ refers to the sigmoid function. The 1×1 convolutional layers $W_1^{1\times1}$ and $W_2^{1\times1}$ have the reduction ratio of r_1 and $\frac{C}{2r_1}$, respectively, stepwise down-sampling the channel dimensions until a size of 2 is reached. Subsequently, the binary mask $M_l \in \{0,1\}^{H\times W}$ for extracting low frequencies can be obtained from the formula $M_l\left[\frac{H}{2}-\alpha\frac{H}{k}:\frac{H}{2}+\alpha\frac{H}{k},\frac{W}{2}-\beta\frac{W}{k}:\frac{W}{2}+\beta\frac{W}{k}\right]=1$, where k is set to 128. Then, set the value of the remaining area to 1 and obtain the masks of the high-frequency M_h.

2.1.2. Frequency Feature Decoupling and Multi-Convolution Cross-Attention Fusion

In this study, the learned masks are applied to the spectrum through element-wise multiplication to isolate specific frequency components. Subsequently, the Inverse Fast Fourier Transform (IFFT) operation is employed to reconvert these masked spectral representations back into the spatial domain. This transformation is essential because the subsequent feature interaction relies on spatial correspondence, ensuring that the decoupled features are dimensionally compatible with the intermediate feature X. After obtaining the adaptive decoupling features, we use the multi-convolution transposed cross-attention (Figure 2c) to excavate different feature parts from the input features under the guidance of low-frequency features F_l and high-frequency features F_h. The feature extraction procedure can be mathematically formulated as follows:

$$\begin{cases} \mathbf{X}_* = softmax\left(\mathbf{Q}\mathbf{K}^\top/\alpha\right)\mathbf{V} \\ \mathbf{Q}=DW_1\left(W_3^{1\times1}(\mathbf{F}_*)\right), \mathbf{K}=DW_2\left(W_4^{1\times1}(\mathbf{X})\right), \mathbf{V}=DW_3\left(W_5^{1\times1}(\mathbf{X})\right) \\ \mathbf{F}_* = \mathcal{F}^{-1}(\mathbf{M}_* \odot \mathbf{F}) \end{cases} \tag{2}$$

where $* \in \{l,h\}$ is a low-frequency or high-frequency index. DW represents a 3×3 convolution (the number in the lower right corner is represented as the number of convolutions of the module). $\odot$ represents element dot multiplication. $\mathcal{F}^{-1}$ is the inverse fast Fourier transform. $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ are queries, keys, and values, respectively. α is a kind of learnable scaling factors used to control the magnitude of the dot multiplication results of $\mathbf{Q}$ and $\mathbf{K}$. *softmax*, the normalization function, calculates $\left(\mathbf{Q}\mathbf{K}^\top/\alpha\right)$ and maps the results to the interval [0,1]. $W_3^{1\times1}$, $W_4^{1\times1}$, and $W_5^{1\times1}$ are 1×1 convolutions. $\mathbf{X}$ represents the original features input to the multi-convolution cross-attention fusion module. $\mathbf{X}_*$ denotes the final output feature after fusion by the multi-convolution cross-attention (including results under low-frequency or high-frequency scenarios), and it is the new feature representation obtained after the attention mechanism acts on the features. $\mathbf{M}_*$ and $*$ correspond to (low-frequency or high-frequency) masks. $\mathbf{F}$ is the fundamental feature, which is the original input value of the masking operation and subsequent attention calculation. $\mathbf{F}_*$ is a low-frequency or high-frequency feature obtained by frequency decoupling.

2.2. Cross-Frequency Domain Feature Reconstruction Module (CFRM)

In this study, a CFRM, as illustrated in Figure 3, is designed through the fusion of optical and SAR imagery. This module facilitates the cross-guidance and reconstruction between low-frequency and high-frequency features, which achieves effective cloud removal of optical imagery. The core goal of the CFRM is to utilize the advantages of different frequency information to complement and guide each other, so as to reduce the impact of cloud occlusion on image detail loss and global information blurring. Specifically, high-frequency features contain rich edge and texture details, and can particularly preserve the structural information of cloud-obscured regions in SAR images. Therefore, this study uses high-frequency features to guide the reconstruction of low-frequency features, aiming at enhancing the detail restoration effect of optical images. This process is realized under the

guidance of an ultra-lightweight spatial attention high-frequency unit for the low-frequency unit (H-L), as shown in Figure 4. Meanwhile, to ensure that the cloud-removed optical image retains the overall hue and shape information, this study leverages the global information embedded in low-frequency features to guide the high-frequency unit (L-H). This information is then transmitted to the high-frequency branch, thereby further optimizing the representation of high-frequency features (see Figure 5). This cross-frequency domain feature complementation mechanism fully leverages the structural and detail advantages of SAR images, integrates the color and texture information of optical images, and thereby effectively enhances the cloud removal performance in both spatial and channel dimensions. Consequently, high-quality optical images can be stably acquired even under heavy cloud occlusion.

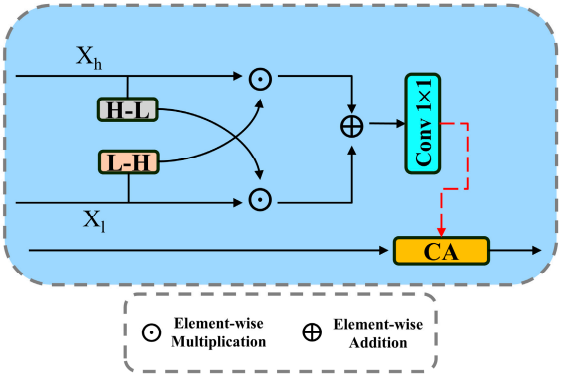

Figure 3. Cross-frequency domain feature reconstruction module.

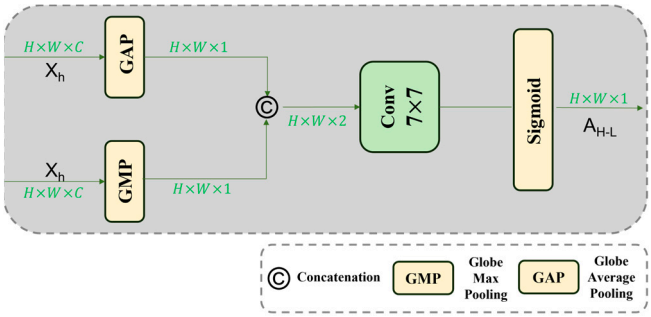

Figure 4. H-L unit.

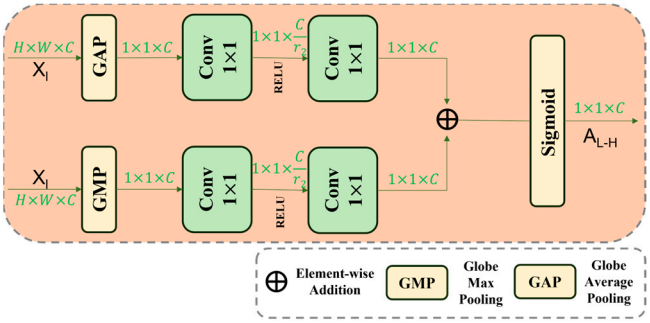

Figure 5. L-H unit.

2.2.1. The High-Frequency Guided Low-Frequency Unit

The high-frequency guided low-frequency unit is illustrated in Figure 4. This unit computes the spatial attention map using features extracted from high-frequency components, which is then utilized to supplement the features in the low-frequency branch. The H-L unit utilizes two different channel pooling techniques to generate two single-channel

spatial feature maps in parallel, each channel being $H \times W \times 1$ in size. These maps are then connected along the channel dimensions. The connected features are further refined by the 7×7 convolutions, and then a sigmoid operation is applied to produce the final spatial attention map. The final step involves obtaining the modulated low-frequency features through an element-wise multiplication.

$$\begin{cases} \hat{\mathbf{X}}_l = \mathbf{X}_l \odot \mathbf{A}_{H-L} \\ \mathbf{A}_{H-L} = \delta\left(W_6^{7\times7}([GAP_c(\mathbf{X}_h), GMP_c(\mathbf{X}_h)])\right) \end{cases} \tag{3}$$

where the channel reduction ratio of the 7×7 convolution W_6 is 2. δ refers to the sigmoid function. GAP_c and GMP_c are the channel-oriented global average pooling and the maximum pooling, respectively. $[\cdots, \cdots]$ represents the join operations. $\hat{\mathbf{X}}_l$ denotes the modulated low-frequency feature, which serves as the final output low-frequency feature processed by the high-frequency guided low-frequency unit. $\mathbf{X}_l$ is the original low-frequency branch feature, which is also the low-frequency input feature to be guided and supplemented by the high-frequency feature. $\mathbf{A}_{H-L}$ represents the spatial attention map generated by the high-frequency guided low-frequency unit. $\mathbf{X}_h$ is the feature extracted from the high-frequency branch, serving as the source of high-frequency information for guiding low-frequency features. The spatial attention map is generated based on this feature, which is then used to supplement information for the low-frequency features.

2.2.2. The Low-Frequency Guided High-Frequency Unit

The low-frequency guided high-frequency unit is illustrated in Figure 5. This unit is a dual-branch module that processes the input low-frequency features and produces a feature descriptor, which is subsequently used to process the high-frequency features. To process the low-frequency feature $\mathbf{X}$, the top branch of the L-H unit leverages global average pooling, which aggregates spatial information to produce a $1 \times 1 \times C$ feature vector. This vector subsequently undergoes processing by two 1×1 convolutional layers, with a ReLU activation function sandwiched between them. The bottom branch of the L-H unit mirrors this architecture, with the only difference being the maximum pooling of the head. Based on the additive result of the two branches, the sigmoid activation is used to generate the final attention map $\mathbf{A}_{L-H} \in R^{1\times1\times C}$, which is then employed to modulate the extracted high-frequency features $\mathbf{X}_h$. The L-H process is formulated as

$$\begin{cases} \hat{\mathbf{X}}_h = \mathbf{X}_h \odot \mathbf{A}_{L-H} \\ \mathbf{A}_{L-H} = \delta\left(W_8^{1\times1}\left(\gamma\left(W_7^{1\times1}(GAP_s(\mathbf{X}_l))\right)\right) + W_{10}^{1\times1}\left(\gamma\left(W_9^{1\times1}(GMP_s(\mathbf{X}_l))\right)\right)\right) \end{cases} \tag{4}$$

where $\hat{\mathbf{X}}_h$ is the modulated high-frequency feature, which is the output result of high-frequency features modulated by the attention descriptor after processing by the low-frequency-guided high-frequency unit, and is intended for subsequent further processing or feature fusion. $\mathbf{X}_h$ represents the original high-frequency feature, which is the high-frequency input feature to be guided and modulated by the low-frequency feature. GAP_s and GMP_s represent global average pooling and spatial max pooling operations, respectively. γ represents the ReLU activation function. $W_7^{1\times1}$, $W_8^{1\times1}$, $W_9^{1\times1}$ and $W_{10}^{1\times1}$ are all 1×1 convolutions. The reduction ratio of adjustment for channels $W_7^{1\times1}$ and $W_9^{1\times1}$ is 2, while the increase ratio of adjustment for channels $W_8^{1\times1}$ and $W_{10}^{1\times1}$ is 2. $\mathbf{A}_{L-H}$ is the spatial attention map generated by the low-frequency guided high-frequency unit. $\mathbf{X}_l$ denotes the feature of the low-frequency branch, which serves as the source of low-frequency information for guiding high-frequency features.

Finally, the modulated high-frequency feature $\hat{\mathbf{X}}_h$ and low-frequency feature $\hat{\mathbf{X}}_l$ are aggregated via the 1×1 convolution to yield the feature $\mathbf{X}_m$, and then the obtained feature

is merged into the original input feature **X** via the cross-attention unit, where feature **X** generates the query **Q** and feature $\mathbf{X}_m$ generates the key **K** and the value **V**. CFRM adaptively adjusts the low-frequency and high-frequency components present in the input feature and realizes adaptive cloud removal according to the degradation types existing in the damaged input images.

3. Experimental Results

3.1. Experiment Design

3.1.1. Experimental Data

The M3R-CR dataset [31] is adopted in this experiment, which contains 63,000 sets of corresponding cloud-free Planetscope multi-spectral images, Sentinel-1 dual-polarization SAR images, and cloudy Planetscope multi-spectral images. Among these images, the Planetscope imagery comprises four spectral bands: blue, green, red, and near-infrared (NIR). The average time interval between the paired cloudy and cloudless data is 2.8 days, and the average time interval between the cloudy data and its corresponding SAR data is 2.2 days.

3.1.2. Methods of Comparison

To verify the effectiveness of the proposed AFR-CR method, this study compares the proposed method with eight advanced cloud removal methods. The cloud removal methods using only optical data include: the cGAN-based cloud removal method (Mc-GAN) [9], the spatial attention-based cloud removal method (SPA-GAN) [39], and the cloud removal method that combines the advantages of CNN and Transformer (MRF-Net) [40]. The cloud removal methods based on SAR and optical data fusion include: the cGAN method with concatenated optical and SAR inputs (SAR-OPT) [28], the cloud removal method based on a deep residual network (DSen2-CR) [29], the SAR-enhanced global-local optical cloud removal method (GLF-CR) [30], the fusion cloud removal method based on optical and SAR pixel alignment (Align-CR) [31], and the multi-layer generative adversarial network based on SAR-to-optical image conversion and cloud removal (MT_GAN) [41].

3.1.3. Experimental Environment and Parameter Settings

The experimental environment was configured as follows: the operating system was Windows 10; the central processing unit (CPU) was an Intel (R) Core (TM) i9-12900K running at a base clock speed of 3.20 GHz; the graphics processing unit (GPU) was an NVIDIA RTX A5000 with 24 GB of video random-access memory; the Compute Unified Device Architecture (CUDA) toolkit version was 11.8; the deep learning framework employed was PyTorch 2.0.0; and the programming language used was Python 3.9. The proposed AFR-CR method was developed using the open-source PyTorch framework and trained end-to-end under L1 loss supervision, utilizing a single NVIDIA A5000 GPU for computation. During the model training phase, the samples in this experiment were randomly cropped into patches of 128 $\times$ 128 pixels. A batch size of 24 was utilized, and the model was trained for a maximum of 30 iterations. Optimization was performed using the Adam algorithm, with an initial learning rate of 0.0001 that was halved every 5 epochs. The division ratio for the training set, validation set, and test set is 8:1:1.

3.1.4. Quantitative Assessment Indicators

The cloud removal performance was quantitatively assessed using four widely adopted metrics in the field, namely the PSNR [35], SSIM [36], MAE [37], and SAM [38]. The pixel-level quality of the cloud-removed imagery was evaluated using PSNR and MAE. SSIM was employed to characterize the restoration of spatial structures, which is grounded

in the principles of visual perception; SAM was used to quantify the spectral information retained in the cloud-removed reconstruction results. Higher PSNR and SSIM values, coupled with lower MAE and SAM values, indicate better quality of the reconstructed images.

3.2. Experimental Results and Analysis

3.2.1. Qualitative Evaluation

The cloud removal results of the proposed method and eight comparative methods under three distinct scenarios are presented in Figure 6. Taking the local thick-cloud coverage in Scenario 1 as an example, notable differences exist in the performance of each method in terms of thick cloud removal and information recovery. Although the SPA-GAN method exhibits a certain degree of cloud removal capability, it still leaves noticeable artifacts in thick cloud regions and introduces overall blurring in the image, leading to severe loss of ground object details. Compared with SPA-GAN, the Mc-GAN method achieves better performance in information recovery for locally thick-cloud-covered areas, particularly in partially reconstructing the obscured ground objects in the upper-left thick cloud region. However, the overall restoration remains incomplete, with some information missing. The SAR-OPT method performs slightly better than Mc-GAN in texture restoration, yet it still fails to effectively recover information in thick-cloud-covered areas. The MRF-Net method performs poorly in thick cloud regions, as it erroneously introduces false ground object information, significantly compromising the reliability of the results. The MT_GAN method shows obvious deficiencies in detail recovery, especially in reconstructing the structural information of buildings obscured by thick clouds. Although the DSen2-CR method successfully removes cloud-covered regions, it exhibits significant deviations in semantic recovery within thick cloud areas, resulting in confusion of ground object categories. The cloud removal results generated by the GLF-CR method contain substantial noise, leading to severe degradation of image quality, particularly in thick cloud regions. The Align-CR method demonstrates good performance in restoring edges in thin cloud regions; however, its restoration effect in local thick cloud areas is limited, showing unnatural transitions and errors in ground object structures. In contrast, the proposed method achieves the best performance under the local thick-cloud conditions of Scenario 1. It not only effectively preserves the spectral characteristics of cloud-free regions, ensuring high spectral consistency between the processed output and the cloud-free reference image, but also more accurately recovers ground object information and edge structures in thick-cloud-covered areas. In particular, the restoration of building contours and surface details is closest to the real cloud-free reference.

In the experimental results of Scenario 2, which involves large-area thick cloud coverage, the performance of each method shows significant disparities when addressing the challenge of severe cloud occlusion. The SPA-GAN method is able to roughly restore the contours of ground objects and demonstrates a certain capability for structural recovery in thick cloud-covered areas; however, it exhibits notable spectral distortion—particularly in coastal regions with thick clouds, where the restoration results deviate considerably from the actual ground features. The Mc-GAN method achieves relatively natural spectral reconstruction in some thick cloud regions, yet cloud removal is incomplete overall, and residual cloud shadows affect both visual quality and usability. The SAR-OPT method performs poorly in restoring images under extensive thick clouds: not only does it fail to remove clouds effectively, but it also leads to severe color distortion in vegetation areas. While the MRF-Net method enhances the edges of ground objects, it fails to effectively remove thick clouds, such as those in the lower-right corner of the image, showing limited capability in recovering information under dense cloud cover. The MT_GAN method is almost unable to achieve effective reconstruction in cloud-obscured regions, especially

over water bodies, making it the weakest performer under large-area thick cloud conditions. The DSen2-CR method reveals clear limitations when dealing with continuous thick clouds, with halo effects and scattering artifacts significantly degrading image quality. The cloud removal results generated by the GLF-CR method contain substantial noise, and overexposure artifacts occur in thick cloud-covered coastal areas, resulting in severe loss of ground object details. Although the Align-CR method performs reasonably well in local spectral restoration, it falls short in recovering detailed textures in thick cloud regions, making it difficult to support fine-grained ground object identification. In contrast, the proposed method still maintains excellent performance in large-area thick cloud scenarios: it effectively recovers water bodies and vegetation information obscured by thick clouds, producing results with clear ground object contours and consistent spectra. This superior restoration is particularly evident in coastal areas with thick clouds, demonstrating strong potential for supporting downstream tasks such as land cover classification.

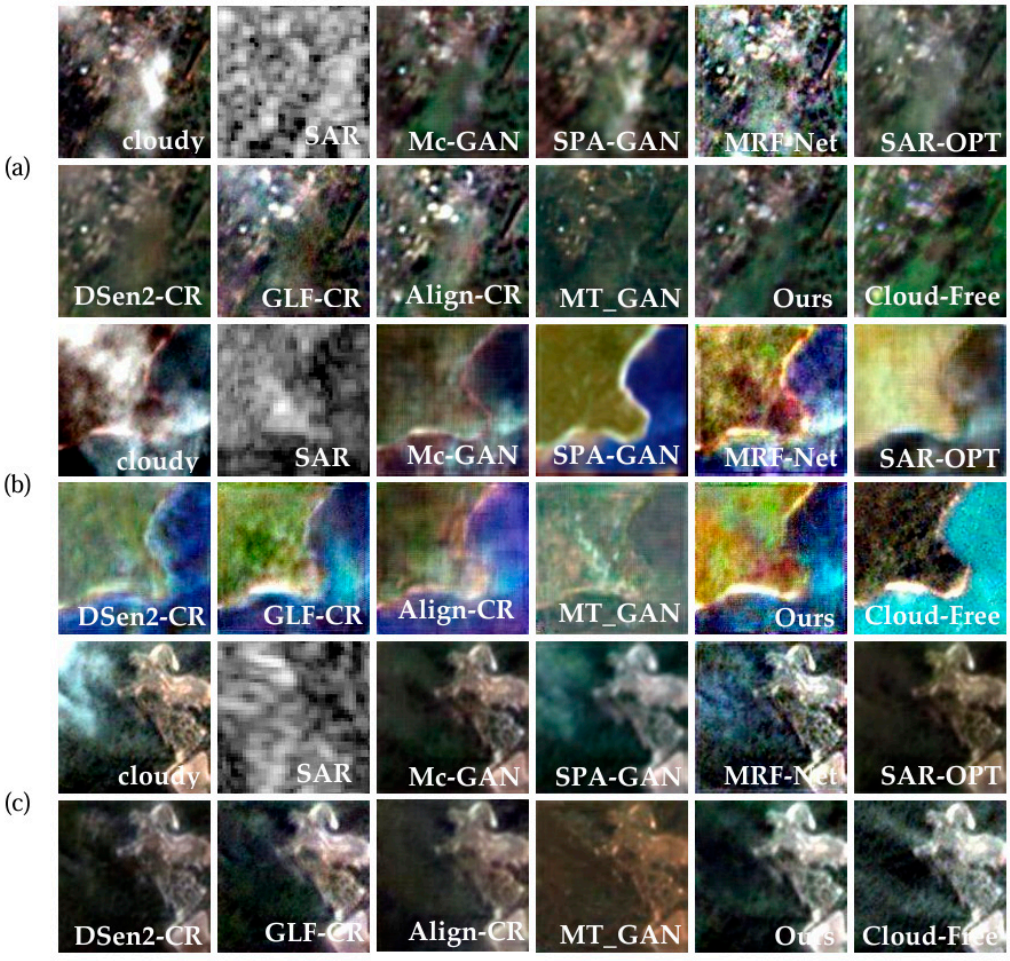

Figure 6. Experimental results of the M3R-CR dataset. (**a**) Scenario1. (**b**) Scenario 2. (**c**) Scenario 3.

In the experimental results of Scenario 3 under thin cloud coverage, the performance of each comparative method in dealing with thin cloud interference still exhibits notable differences. The SPA-GAN method continues to suffer from a certain degree of image blurring, with limited capability in restoring terrain contours and details, resulting in unclear classification boundaries. The Mc-GAN method remains unable to adequately recover fine textures in mountainous areas under thin cloud coverage, leading to some loss of high-frequency information. Although the SAR-OPT and MT_GAN methods improve texture depiction in certain areas by incorporating SAR data, they still fall short in terms of semantic consistency and object discrimination across the entire scene. The

MRF-Net method achieves partial cloud removal to some extent but shows instability in spectral preservation, with localized spectral distortion occurring. The DSen2-CR method demonstrates excellent spectral fidelity and detail restoration in cloud-free regions; however, its performance in reconstructing thin cloud-covered areas remains relatively limited, particularly in restoring semantic information and the structural details of ground objects beneath the clouds. Both the GLF-CR and Align-CR methods also show certain limitations in thin cloud scenarios: GLF-CR is still affected by artifacts and halo effects at edges, while Align-CR is less accurate than the proposed method in reconstructing fine contour details of ground objects.

To summarize, the cloud removal method proposed in this study, which integrates spatial-domain and frequency-domain characteristics, has demonstrated superior performance across multiple scenarios. By incorporating the SAR auxiliary information, the method achieves a significant improvement in both cloud removal accuracy and image restoration quality. Evaluated alongside eight leading-edge methods for optical image cloud removal, the proposed method demonstrates greater effectiveness in addressing complex cloud-occluded regions and enables the restoration of more abundant ground object details and semantic information.

3.2.2. Quantitative Evaluation

Table 1 displays the quantitative evaluation results of the nine cloud removal methods across all test sets. The experimental results demonstrate the superior performance of our proposed method across all evaluated metrics. Specifically, it yields an SSIM value of 0.918, a PSNR of 31.378 dB, a SAM as low as 5.275, and an MAE of merely 0.023. These numerical findings confirm the superiority of the proposed method over other compared approaches in terms of reconstruction quality and accuracy. The Mc-GAN and SPA-GAN methods belong to image reconstruction approaches based on spectral information. While both methods yield improvements in image quality, with Mc-GAN achieving an SSIM value of 0.903 and a PSNR value of 30.181 dB, they primarily rely on spectral information for the reconstruction process. This inherent dependence imposes certain limitations on their ability to suppress noise and preserve image details. In particular, regarding the two metrics of SAM and MAE, the SAM and MAE values of the Mc-GAN method are 6.047 and 0.026, respectively, whereas those of the SPA-GAN method are 6.416 and 0.028, respectively. The MRF-Net method also has the aforementioned limitations in both detail restoration and noise suppression. Although this type of spectrum-based simple reconstruction method can restore the structural features of images to a certain extent, it often fails to effectively eliminate noise in SAR images, thereby inducing spectral distortion. In comparison, the SAR-OPT method fuses SAR and optical imagery within a generative adversarial network (GAN) framework. This fusion strategy improves the quality of image reconstruction, especially on SSIM and PSNR. However, the fusion strategy employed by this method is relatively simple and fails to fully explore the synergistic advantages of SAR and optical images. Therefore, its SAM only reaches a value of 5.911, indicating that its cloud removal performance is still inadequate. The SAR-OPT method has a favorable performance in restoring low-resolution image information; however, it cannot effectively solve the issues of noise interference and spectral deviation. This limitation is particularly pronounced in complex environmental application scenarios, where unnecessary detail loss tends to occur. Among all the methods, the MT_GAN method exhibits the poorest performance, with a PSNR value of merely 26.745 dB. In this model, issues including SAR geometric deformation, foreshortening, and layover impair the accuracy and reliability of the experiment, which in turn degrades the quality of the optical images generated by the MT_GAN method. The DSen2-CR method adopts the network architecture composed of

multi-residual blocks, which enhances its capability to extract deep-level features. This method improves its cloud removal capability through the design of multi-scale learning and the deep residual blocks. However, due to insufficient consideration of the deep fusion between SAR noise and spectral information, both the spectral consistency and detail restoration performance of the resulting images remain inadequate. Furthermore, the DSen2-CR method is less robust when dealing with high noise environments, resulting in its inferiority to other methods on SAM and MAE. The GLF-CR method incorporates SAR filtering and global context information. Comprehensively accounting for the noise characteristics of SAR images and the spatial context of images, it improves the cloud removal efficacy and structural similarity of the processed images, with corresponding SSIM and PSNR values of 0.896 and 29.908 dB, respectively. Nevertheless, the GLF-CR method still exhibits room for improvement in terms of image detail reconstruction and spectral consistency. Especially when regarding the detail recovery during cloud removal and reconstruction, the GLF-CR method fails to fully exploit the potential of frequency-domain information, resulting in its performance on the SAM and MAE metrics lagging behind that of the proposed method. The Align-CR method solves the geometric mismatch problem between images through a pixel-level alignment strategy, which significantly enhances the structural similarity and the reconstruction quality of images. However, the Align-CR method does not fully leverage the depth information of the frequency domain, which consequently imposes restrictions on its performance in terms of spectral consistency and detail restoration. In particular, with respect to SAR noise elimination and image detail enhancement, the Align-CR method demonstrates marginally inferior performance compared to the proposed method.

Table 1. Quantitative evaluation indicators.

Method	SSIM ↑	PSNR (dB) ↑	SAM ↓	MAE ↓	w/o SAR
Mc-GAN	0.903	30.181	6.047	0.026	×
SPA-GAN	0.893	29.672	6.416	0.028	×
MRF-Net	0.888	29.810	6.350	0.027	×
SAR-OPT	0.906	30.191	5.911	0.026	√
DSen2-CR	0.884	29.363	6.847	0.029	√
GLF-CR	0.896	29.908	5.971	0.027	√
Align-CR	<u>0.910</u>	<u>30.338</u>	<u>5.561</u>	<u>0.025</u>	√
MT_GAN	0.753	26.745	9.676	0.039	√
AFR-CR	**0.918**	**31.378**	**5.275**	**0.023**	√

The best is bolded and the second best is underlined. The abbreviation w/o SAR denotes the model without SAR data input. ("↑" indicates that larger values are better, "↓" indicates that smaller values are better, "×" indicates that the original model is not a multimodal fusion method for SAR, and "√" indicates that the original model is a multimodal fusion method for SAR.).

The cloud removal method proposed in this study outperforms the existing comparative methods across all evaluation metrics. Its core advantage lies in its effective fusion of multispectral and SAR imagery for cloud removal. By strategically introducing frequency-domain information, the method simultaneously mitigates SAR-specific noise in high frequencies and recovers cloud-corrupted low-frequency details, leading to significant improvements across all evaluation metrics.

3.3. Ablation Study

3.3.1. Ablation Experiments of the AFR-CR Method on the M3R-CR Dataset

The AFR-CR method proposed in this study effectively extracts and separates the low-frequency and high-frequency information of images by introducing the AFRB module. Specifically, it restores missing information and preserves the global consistency of

images through cross-guided reconstruction between low-frequency and high-frequency features, thereby enhancing the cloud removal effect achieved through the fusion assisted by synthetic aperture radar data. To elucidate the factors for the enhanced performance of the proposed approach, the contribution of each component was assessed by comparing two versions of the model: one excluding SAR imagery and the other omitting the AFRB module. The quantitative and qualitative results are presented in Table 2 and Figure 7.

Table 2. Quantitative evaluation indicators of the ablation experiment.

Method	SSIM ↑	PSNR (dB) ↑	SAM ↓	MAE ↓	w/o SAR
w/o AFRB	<u>0.906</u>	<u>30.558</u>	<u>6.010</u>	<u>0.025</u>	√
w/o SAR	0.893	30.093	6.618	0.027	×
AFR-CR	**0.918**	**31.378**	**5.275**	**0.023**	√

The best is bolded and the second best is underlined. The abbreviation w/o SAR denotes the model without SAR data input. The abbreviation w/o AFRB indicates that the model excludes the AFRB. ("↑" indicates that larger values are better, "↓" indicates that smaller values are better, "×" indicates no SAR data input, "√" indicates SAR data input).

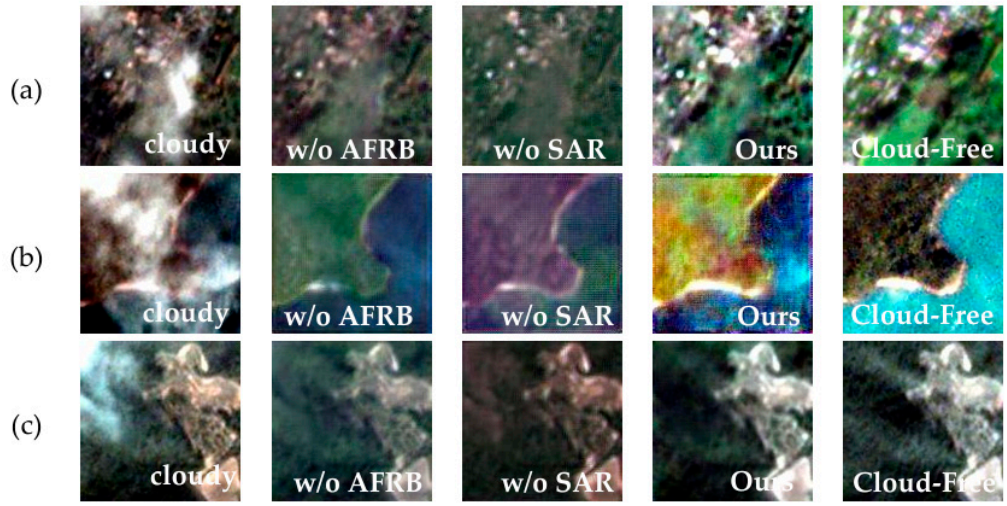

Figure 7. Qualitative results of the ablation experiment. (**a**) Scenario1. (**b**) Scenario 2. (**c**) Scenario 3.

It can be observed from Table 2 that the method proposed in this study shows the best performance across all evaluation indicators. Specifically, in terms of PSNR, the method proposed in this study reaches a value of 31.378, which is 0.820 and 1.285 higher than those of w/o AFLB and w/o SAR variants, respectively. This result indicates that the proposed method preserves more image details during the cloud removal process. In addition, the proposed method has particularly favorable performance in terms of SAM. Compared with the w/o AFLB and w/o SAR variants, its SAM value is improved by 0.735 and 1.343, respectively, which corresponds to a smaller spectral error. This outcome indicates that the cloud-removed reconstructed image is spectrally more consistent with the real image, thereby mitigating spectral distortion during the cloud removal process. In terms of SSIM, the proposed method also outperforms the other two comparative methods, which indicates that it has higher fidelity in visual structure similarity and can better preserve the structural information of the original image. Finally, in terms of the MAE metric, the proposed method exhibits a lower MAE value than the w/o AFRB and w/o SAR variants. This result indicates that the proposed method produces the minimum error during the cloud removal process, thereby yielding more accurate reconstructed images.

It is noteworthy that in the absence of SAR data assistance, the w/o SAR variant exhibits generally poor performance across all evaluation metrics, with particularly notable deficiencies in the SAM and MAE metrics. This phenomenon can be attributed to the lack

of complementary information afforded by SAR data, which ultimately leads to inadequate spectral recovery capabilities and compromised error suppression efficiency. Although the w/o AFRB method exhibits suboptimal performance across the four indicators, its overall performance is marginally inferior to that of the AFR-CR method. This discrepancy can be attributed to the relative simplicity of its model architecture: while the method is capable of preserving a certain degree of structural fidelity, it lacks the adaptive frequency modulation mechanism inherent to the AFRB module, thereby restricting the detail restoration capability of its reconstruction results in complex scenarios. In conclusion, by integrating SAR data assistance and the AFRB module, the proposed approach demonstrates enhanced capabilities in maintaining structural integrity, ensuring spectral fidelity, and minimizing reconstruction errors, but the fused image also demonstrates superior cloud removal performance compared to the other methods.

The qualitative results of the ablation experiment are shown in Figure 7. As observed from the concentrated cloud regions, the w/o AFRB variant presents varying degrees of scattering artifacts in its reconstruction results. This phenomenon is attributed to the absence of the adaptive frequency reconstruction capability inherent to the AFRB module, which further results in the failure of high-frequency details to fully utilize the effective information during the decoding process. The ground object information restored by the w/o SAR variant lacks details. This is because the variant solely relies on the local contextual information of optical images for interpolation-based inpainting, while lacking the prior structural information of cloud-penetrating ground objects provided by SAR data, ultimately diminishing the credibility of the reconstruction results. In contrast, the proposed method significantly suppresses scattering artifacts by integrating the AFRB module and SAR data-assisted input, rendering the reconstructed image more consistent with the spectral and spatial structures of the reference image.

3.3.2. Impact of Frequency Split Factor

The parameter K serves as the frequency split factor in our AFR module, determining the baseline granularity of the dynamic mask generation. It controls the scaling of the frequency bandwidth before adaptive adjustment. To verify the optimal setting, we conducted a sensitivity analysis by varying K across values of 64, 128, 256, 512, and 1024.

Figure 8 visualizes the trends of four key metrics, including PSNR, SSIM, SAM, and MAE. It is evident that the model performance exhibits a distinct peak at $K = 128$. Specifically, the reconstruction quality metrics achieve their maximums at this setting, reaching a PSNR of 31.378 dB and an SSIM of 0.918. Correspondingly, the error metrics drop to their minimums, recording a SAM of 5.275 and an MAE of 0.023. Increasing or decreasing K leads to observable degradation in all metrics, confirming that $K = 128$ provides the most effective frequency separation granularity.

3.4. The Computational Efficiency of AFR-CR

We also conducted a comprehensive evaluation of computational efficiency for the proposed AFR-CR method alongside multiple comparative approaches. To ensure a fair comparison, we focused on representative methods within the same technical domain (i.e., SAR-assisted cloud removal), as they share similar input modalities and fusion challenges [31,42]. The metrics employed include model parameters (Params), floating-point operations (FLOPs), inference time, and PSNR. The specific results are illustrated in Figure 9. It is noteworthy that although AFR-CR exhibits a slightly higher number of Params than both DSen-CR and GLF-CR, indicating a relatively more complex structure, its FLOPs and inference time remain at comparatively low levels, demonstrating excellent computational efficiency. This advantage primarily stems from the attention mechanism and feature

fusion strategy employed by AFR-CR, which significantly reduces redundant computations while fully preserving critical information. Moreover, AFR-CR maintains a high PSNR metric, indicating that efficiency gains are achieved without compromising reconstruction quality. Overall, AFR-CR strikes a favorable balance between computational complexity and reconstruction performance.

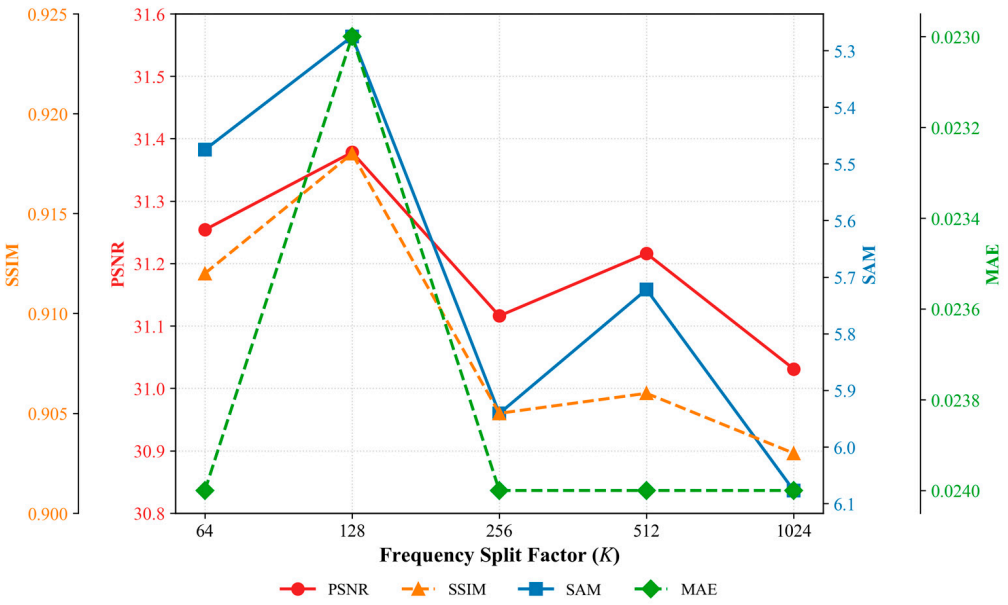

Figure 8. Sensitivity analysis of the frequency split factor *K*.

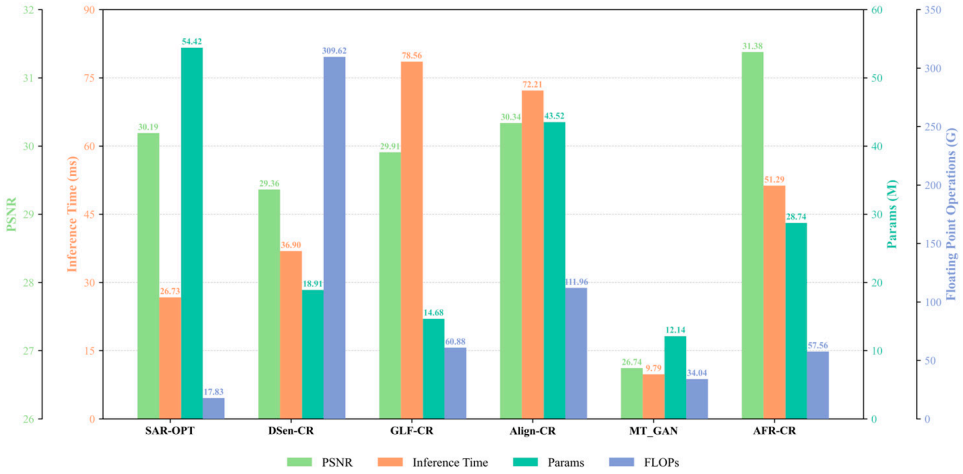

Figure 9. The computational efficiency of the AFR-CR and other selected State-of-the-Art methods.

4. Discussion

This section focuses on examining a critical question: how does its cloud removal performance compare to alternative methods under varying cloud cover conditions? Figure 10 provides a quantitative comparison of image quality trends across different methods within the 0% to 100% cloud cover range. Figure 11 shows the experimental results of ablation under different cloud cover conditions.

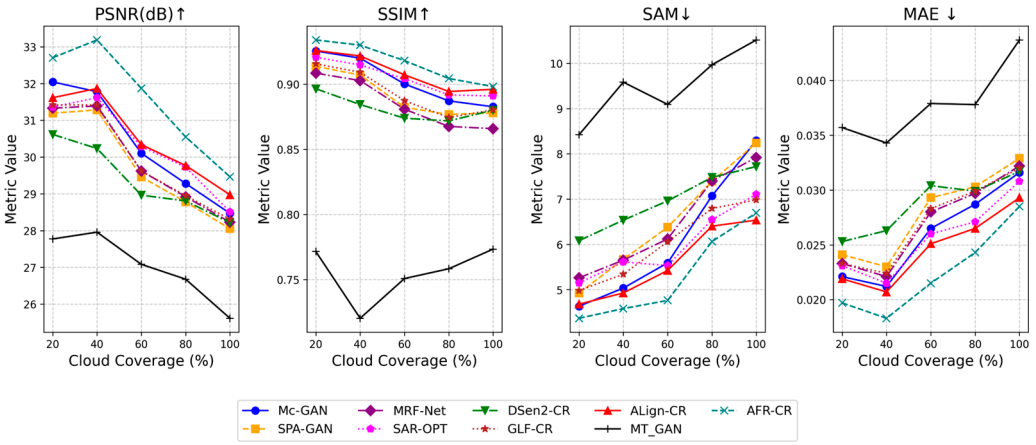

Figure 10. Analysis of different cloud coverage.

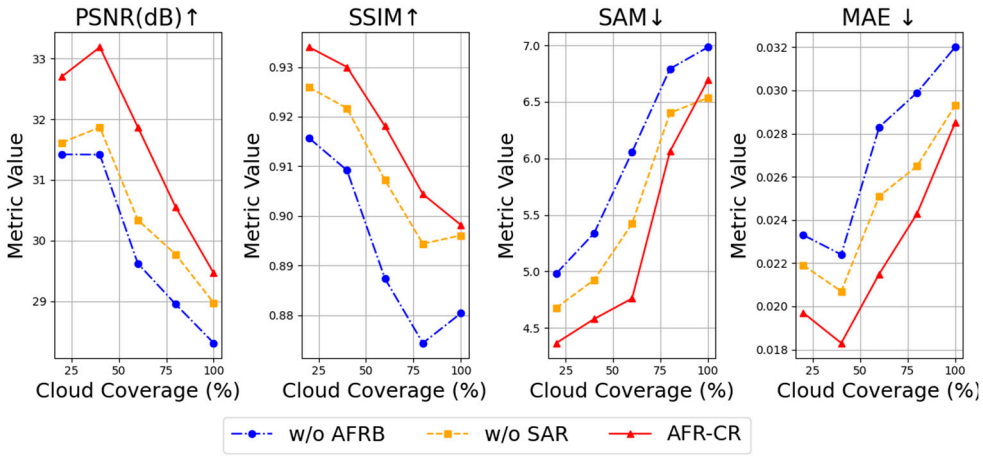

Figure 11. Experimental results of different cloud level ablation based on PSNR, SSIM, SAM, and MAE quality indexes.

4.1. Analysis of Different Cloud Cover Levels on the M3R-CR Dataset

This study further conducts a comparative analysis that evaluates the performance of the proposed AFR-CR method against eight advanced cloud removal methods across different cloud cover levels. Figure 10 illustrates the evaluation results for the four metrics: PSNR, SSIM, SAM, and MAE across five cloud cover intervals: 0% to 20%, 20% to 40%, 40% to 60%, 60% to 80%, and 80% to 100% (corresponding to abscissa values of 20, 40, 60, 80, and 100 in Figure 10). Compared with the eight advanced comparative methods, the AFR-CR method achieves the optimal overall performance across all cloud coverage levels. Figure 10 reveals an overall inverse relationship between cloud coverage level and the performance of all cloud removal methods. Specifically, higher cloud coverage compromises the spectral distinctiveness required by the FDDM for precise decoupling and reduces the valid optical anchors needed by the CFRM for cross-modal alignment. This scarcity of prior information inherently limits the model's ability to fully leverage frequency interactions, resulting in inferior performance. When the cloud coverage ranges from 0% to 20%, the Mc-GAN method has an overall favorable performance. Across the PSNR and SAM metrics, it ranks second, surpassed only by the proposed method in this study; additionally, its SSIM and MAE metrics are comparable to those of the sub-optimal methods. Both the SPA-GAN method and the MRF-Net method perform poorly under different cloud cover conditions. The MT_GAN method shows the worst performance under different cloud covers. The performance of the DSen2-CR method decreases sharply as cloud coverage increases. While the method exhibits a certain level of performance in

terms of basic metrics under low cloud coverage, its accuracy and structural preservation capabilities are significantly compromised under high cloud coverage. The Align-CR, GLF-CR, and SAR-OPT methods show a certain degree of effectiveness under low cloud coverage. When cloud coverage exceeds 40%, their PSNR and SSIM metrics deteriorate rapidly, indicating poor adaptability to high cloud coverage scenarios. In contrast, our method effectively mitigates severe structural degradation by anchoring on SAR-derived high-frequency priors within the CFRM, thereby preserving geometric fidelity even when optical information is scarce. In general, the Align-CR method and the proposed method exhibit prominent advantages across all cloud coverage conditions; among these two methods, the proposed method outperforms the Align-CR method in terms of the PSNR, SSIM, and MAE metrics.

4.2. Analysis of Different Cloud Cover Levels in Ablation Study

The outcomes of the ablation study were qualitatively assessed using quantitative metrics under a range of cloud coverage scenarios and are presented in Figure 11. As observed from the figure, the overall performance of all tested methods exhibits a decreasing trend with an increase in the cloud cover percentage. As cloud cover intensifies, the w/o AFRB variant exhibits a more rapid decline in removal performance compared to the w/o SAR variant. This observation indicates that the AFRB module proposed in this study can mitigate the aforementioned performance degradation trend to a certain extent. The lack of supplementary information provided by SAR data results in a significant degradation of the evaluation metrics for the w/o SAR variant when the cloud coverage exceeds 50%. Among all the ablation study variants, the AFR-CR method achieves the optimal performance. The proposed method, AFR-CR, can more effectively utilize the supplementary information from SAR data by leveraging frequency-domain transformation and the cross-guidance between high-frequency and low-frequency features during the reconstruction process. Its comprehensive performance outperforms that of the w/o AFRB variant and the w/o SAR variant across all cloud coverage levels.

5. Conclusions

Cloud removal in remote sensing imagery plays a pivotal role in advancing the availability of optical images. However, existing SAR-assisted cloud removal approaches universally suffer from a limitation: the feature fusion process fails to adequately integrate frequency-domain features, thereby preventing the simultaneous elimination of both cloud cover and noise. This inherent flaw often manifests as blurred edges and diminished detail in the subsequently reconstructed cloud-free images. As a result, the proposed adaptive frequency domain feature reconstruction-based cloud removal method leverages the SAR imagery to facilitate recovery of cloud-obscured optical imagery, thereby addressing the limitations of conventional approaches in terms of feature fusion and detail preservation capability. The proposed FDDM effectively extracts and separates the low and high-frequency information of images, suppressing the interference of low-frequency components to the clouds while maintaining high-frequency details of the surface. The CFRM further practices the cross-guide reconstruction between low and high-frequency features, leading to a great improvement in the fidelity of the reconstructed imagery. The proposed method was benchmarked against eight contemporary leading methods using the M3R-CR dataset, demonstrating superior performance. In comparison to the second-best method, the proposed approach yields improvements of 0.009, 1.040, 0.286, and 0.002 across the four metrics of SSIM, PSNR, SAM, and MAE, respectively.

The practical significance of this research lies in enhancing the usability of optical imagery under cloud-contaminated conditions, providing crucial support for remote sensing

applications reliant on high-quality optical imagery, such as disaster emergency monitoring, crop classification, and land use/cover identification. Although the current method demonstrates excellent performance in reconstruction quality, there remains scope for optimization in computational efficiency and generalization capability. Future research will prioritize optimizing computational efficiency to facilitate deployment on resource-constrained edge devices. To this end, we plan to investigate lightweight attention mechanisms and model quantization techniques to reduce network complexity without compromising performance. Beyond efficiency improvements, we also aim to extend the practical utility of our approach to downstream tasks, such as post-cloud-removal feature classification and change detection, thereby broadening its application in multimodal remote sensing.

Author Contributions: Conceptualization, X.Z. and Q.F.; methodology, X.Z.; software, Q.F.; validation, X.G. and S.Y.; formal analysis, S.Y.; resources, X.G.; data curation, X.Z., Q.F. and S.Y.; writing—original draft preparation, X.Z., Q.F. and X.G.; writing—review and editing, T.L.; visualization, T.L.; supervision, Y.W., A.M. and Y.Z.; funding acquisition, X.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Open Fund of State Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University under Grant 23R03, the Jiangxi Provincial Natural Science Foundation under Grant 20252BAC240247, and the Funding of National Key Laboratory of Uranium Resources Exploration-Mining and Nuclear Remote Sensing under Grant 2025QZ-YZZ-08.

Data Availability Statement: The M3R-CR dataset is available online at https://gitlab.lrz.de/ai4eo/M3R-CR (accessed on 16 May 2024).

Acknowledgments: The authors would like to thank the researchers from the School of Electronic Information at Wuhan University and the Data Science in Earth Observation group at the Technical University of Munich for making their dataset publicly available, which has been instrumental in this study.

Conflicts of Interest: Author Qirui Fang was employed by the company Nanjing Krystal Vision Technologies. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Yuan, Q.; Shen, H.; Li, T.; Li, Z.; Li, S.; Jiang, Y.; Xu, H.; Tan, W.; Yang, Q.; Wang, J. Deep learning in environmental remote sensing: Achievements and challenges. *Remote Sen. Environ.* **2020**, *241*, 111716.
2. Requena-Mesa, C.; Benson, V.; Reichstein, M.; Runge, J.; Denzler, J. EarthNet2021: A large-scale dataset and challenge for Earth surface forecasting as a guided video prediction task. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtually, 19–25 June 2021; pp. 1132–1142.
3. Shastry, A.; Carter, E.; Coltin, B.; Sleeter, R.; McMichael, S.; Eggleston, J. Mapping floods from remote sensing data and quantifying the effects of surface obstruction by clouds and vegetation. *Remote Sen. Environ.* **2023**, *291*, 113556. [CrossRef]
4. Ma, Y.; Chen, S.; Ermon, S.; Lobell, D.B. Transfer learning in environmental remote sensing. *Remote Sens. Environ.* **2024**, *301*, 113924. [CrossRef]
5. 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]
6. Ebel, P.; Schmitt, M.; Zhu, X.X. Cloud removal in unpaired Sentinel-2 imagery using cycle-consistent GAN and SAR-optical data fusion. In Proceedings of the IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium, Virtually, 26 July–2 August 2020; pp. 2065–2068.
7. Li, J.; Wu, Z.; Hu, Z.; Zhang, J.; Li, M.; Mo, L.; Molinier, M. Thin cloud removal in optical remote sensing images based on generative adversarial networks and physical model of cloud distortion. *ISPRS J. Photogramm. Remote Sens.* **2020**, *166*, 373–389. [CrossRef]
8. Liu, J.; Pan, B.; Shi, Z. Cascaded memory network for optical remote sensing imagery cloud removal. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5613611. [CrossRef]

9. Enomoto, K.; Sakurada, K.; Wang, W.; Fukui, H.; Matsuoka, M.; Nakamura, R.; Kawaguchi, N. Filmy cloud removal on satellite imagery with multispectral conditional generative adversarial nets. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21 July–26 July 2017; pp. 48–56.
10. Singh, P.; Shree, R. Impact of method noise on SAR image despeckling. *Int. J. Inf. Technol. Web* **2020**, *15*, 52–63. [CrossRef]
11. Gong, X.; Hou, Z.; Wan, Y.; Zhong, Y.; Zhang, M.; Lv, K. Multispectral and SAR image fusion for multiscale decomposition based on least squares optimization rolling guidance filtering. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5401920. [CrossRef]
12. Gong, X.; Luo, Y.; Chen, W.; Chang, Y.; Wan, Y.; Ma, A.; Zhong, Y. BASHVS: A Multispectral and SAR Image Fusion Method Based on Bidirectional Aggregation of Saliency in Human Visual System. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5405615. [CrossRef]
13. Li, X.; Wang, L.; Cheng, Q.; Wu, P.; Gan, W.; Fang, L. Cloud removal in remote sensing images using nonnegative matrix factorization and error correction. *ISPRS J. Photogramm. Remote Sens.* **2019**, *148*, 103–113. [CrossRef]
14. Singh, P.; Shree, R.; Diwakar, M. A new SAR image despeckling using correlation based fusion and method noise thresholding. *J. King Saud Univ.-Comput. Inf. Sci.* **2021**, *33*, 313–328. [CrossRef]
15. Chan, T.F.; Shen, J. Nontexture inpainting by curvature-driven diffusions. *J. Vis. Commun. Image Represent.* **2001**, *12*, 436–449. [CrossRef]
16. 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]
17. Yu, Z.; Idris, M.Y.I.; Wang, P. DC4CR: When Cloud Removal Meets Diffusion Control in Remote Sensing. *arXiv* **2025**, arXiv:2504.14785.
18. 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]
19. Ma, X.; Huang, Y.; Zhang, X.; Pun, M.-O.; Huang, B. Cloud-egan: Rethinking cyclegan from a feature enhancement perspective for cloud removal by combining cnn and transformer. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2023**, *16*, 4999–5012. [CrossRef]
20. Yu, X.; Pan, J.; Xu, J.; Wang, M. Missing information reconstruction integrating isophote constraint and color-structure control for remote sensing data. *ISPRS J. Photogramm. Remote Sens.* **2024**, *208*, 261–278. [CrossRef]
21. 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]
22. Ebel, P.; Xu, Y.; Schmitt, M.; Zhu, X.X. SEN12MS-CR-TS: A remote-sensing data set for multimodal multitemporal cloud removal. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5222414. [CrossRef]
23. 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]
24. Shen, H.; Wu, J.; Cheng, Q.; Aihemaiti, M.; Zhang, C.; Li, Z. A spatiotemporal fusion based cloud removal method for remote sensing images with land cover changes. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2019**, *12*, 862–874. [CrossRef]
25. Li, M.; Xu, Q.; Li, K.; Li, W. DecloudFormer: Quest the key to consistent thin cloud removal of wide-swath multi-spectral images. *Pattern Recognit.* **2025**, *166*, 111664. [CrossRef]
26. Li, X.; Zhao, X.; Wang, F.; Ren, P. HF-T2CR: High-fidelity thin and thick cloud removal in optical satellite images through SAR fusion. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4105813. [CrossRef]
27. Li, C.; Li, S.; Liu, X. Breaking through clouds: A hierarchical fusion network empowered by dual-domain cross-modality interactive attention for cloud-free image reconstruction. *Inform. Fusion* **2025**, *113*, 102649. [CrossRef]
28. Grohnfeldt, C.; Schmitt, M.; Zhu, X. A conditional generative adversarial network to fuse SAR and multispectral optical data for cloud removal from Sentinel-2 images. In Proceedings of the IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018; pp. 1726–1729.
29. 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]
30. 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]
31. Xu, F.; Shi, Y.; Ebel, P.; Yang, W.; Zhu, X.X. Multimodal and Multiresolution Data Fusion for High-Resolution Cloud Removal: A Novel Baseline and Benchmark. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5600615. [CrossRef]
32. Zhang, M.; Zhang, C.; Zhang, Q.; Guo, J.; Gao, X.; Zhang, J. ESSAformer: Efficient transformer for hyperspectral image super-resolution. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 23073–23084.
33. Zhang, M.; Bai, H.; Zhang, J.; Zhang, R.; Wang, C.; Guo, J.; Gao, X. RKformer: Runge-Kutta Transformer with Random-Connection Attention for Infrared Small Target Detection. In Proceedings of the 30th ACM International Conference on Multimedia, Lisboa, Portugal, 10–14 October 2022; pp. 1730–1738.

34. Liu, J.; Pan, B.; Shi, Z. CR-Famba: A frequency-domain assisted mamba for thin cloud removal in optical remote sensing imagery. *IEEE Trans. Multimed.* **2025**, *27*, 5659–5668. [CrossRef]
35. Ajayi, O.G.; Ojima, A. Performance evaluation of selected cloud occlusion removal algorithms on remote sensing imagery. *Remote Sens. Appl.* **2022**, *25*, 100700. [CrossRef]
36. Bakurov, I.; Buzzelli, M.; Schettini, R.; Castelli, M.; Vanneschi, L. Structural similarity index (SSIM) revisited: A data-driven approach. *Expert Syst. Appl.* **2022**, *189*, 116087. [CrossRef]
37. Hodson, T.O. Root mean square error (RMSE) or mean absolute error (MAE): When to use them or not. *Geosci. Model Dev.* **2022**, *2022*, 5481–5487. [CrossRef]
38. Azam, M.A.; Khan, K.B.; Salahuddin, S.; Rehman, E.; Khan, S.A.; Khan, M.A.; Kadry, S.; Gandomi, A.H. A review on multimodal medical image fusion: Compendious analysis of medical modalities, multimodal databases, fusion techniques and quality metrics. *Comput. Biol. Med.* **2022**, *144*, 105253. [CrossRef] [PubMed]
39. Sarukkai, V.; Jain, A.; Uzkent, B.; Ermon, S. Cloud removal from satellite images using spatiotemporal generator networks. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Snowmass Village, CO, USA, 1–5 March 2020; pp. 1796–1805.
40. Xu, Q.; Chen, J.; Yan, X.; Li, W. MRF-Net: An infrared remote sensing image thin cloud removal method with the intra-inter coherent constraint. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5643819. [CrossRef]
41. Wang, P.; Chen, Y.; Huang, B.; Zhu, D.; Lu, T.; Dalla Mura, M.; Chanussot, J. MT_GAN: A SAR-to-optical image translation method for cloud removal. *ISPRS J. Photogramm. Remote Sens.* **2025**, *225*, 180–195. [CrossRef]
42. Pan, J.; Xu, J.; Yu, X.; Ye, G.; Wang, M.; Chen, Y.; Ma, J. HDRSA-Net: Hybrid dynamic residual self-attention network for SAR-assisted optical image cloud and shadow removal. *ISPRS J. Photogramm. Remote Sens.* **2024**, *218*, 258–275. [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.