remote sensing

MDPI

Article

A Multi-Level SAR-Guided Contextual Attention Network for Satellite Images Cloud Removal

Ganchao Liu, Jiawei Qiu and Yuan Yuan *

School of Artificial Intelligence, OPtics and ElectroNics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, China; liuganchao@nwpu.edu.cn (G.L.); qiujiawei@mail.nwpu.edu.cn (J.Q.)
* Correspondence: y.yuan1.ieee@gmail.com

Abstract: In the field of remote sensing, cloud cover severely reduces the quality of satellite observations of the earth. Due to the complete absence of information in cloud-covered regions, cloud removal with a single optical image is an ill-posed problem. Since the synthetic aperture radar (SAR) can effectively penetrate clouds, fusing SAR and optical remote sensing images will effectively alleviate this problem. However, existing SAR-based optical cloud removal methods fail to effectively leverage the global information provided by the SAR image, resulting in limited performance gains. In this paper, we introduce a novel cloud removal method named the Multi-Level SAR-Guided Contextual Attention Network (MSGCA-Net). MSGCA-Net is designed with a multi-level architecture that integrates a SAR-Guided Contextual Attention (SGCA) module to fuse the dependable global contextual information from SAR images with the local features of optical images effectively. In the module of SGCA, the SAR image provides reliable global contextual information and genuine structure of cloud-covered regions, while the optical image provides the local feature information. The proposed model can efficiently extract and fuse global and local contextual information in SAR and optical images. We trained and evaluated the performance of the model on both simulated and real-world datasets. Both qualitative and quantitative experimental evaluation demonstrated that the proposed method can yield high quality cloud-free images and outperform state-of-the-art cloud removal methods.

Keywords: SAR; cloud removal; multi-source fusion; image restoration

Citation: Liu, G.; Qiu, J.; Yuan, Y. A Multi-Level SAR-Guided Contextual Attention Network for Satellite Images Cloud Removal. *Remote Sens.* **2024**, *16*, 4767. https://doi.org/10.3390/rs16244767

Academic Editor: Dusan Gleich

Received: 17 October 2024
Revised: 7 December 2024
Accepted: 13 December 2024
Published: 20 December 2024

1. Introduction

Earth observation through remote sensing satellites plays a crucial role in monitoring information in various fields, such as object detection [1], scene classification [2], visual localization [3], and land cover change detection [4]. Significant advancements in optical remote sensing satellites have contributed to these tasks. China has made notable contributions with its high-resolution satellite fleet, including the Jilin-1, Zhuhai-1, Environment-1 (HJ-1), and the Gaofen (GF) series, which comprises 14 satellites. These advancements have brought about an abundance of precise, diverse, and highly time-sensitive remote sensing data. Although the quality and quantity of remote sensing satellite observation images have been increased dramatically in recent years, the process of acquiring optical images is inevitably interfered with by clouds in the atmosphere.

As thick clouds are opaque across all optical bands, they completely obscure the reflected signal, thereby blocking observation of the Earth's surface. In contrast, thin clouds and cloud shadows do not entirely obstruct ground object information, but can still significantly degrade image quality and introduce artifacts. A study [5] by the Moderate Resolution Imaging Spectroradiometer (MODIS) shows that the overall global cloud cover is about 67%, and the cloud cover over land is about 55%. Extensive cloud cover significantly diminishes the quality of Earth observation data, thereby compromising the subsequent processing and applications of imagery. For example, for agricultural monitoring, consistent

time-series data are required, and for disaster monitoring, observations of a particular scene at a particular time are required. Therefore, reconstructing remote sensing images with cloud cover will significantly improve the availability of remote sensing data with major economic and social benefits.

In general, cloud removal in remote sensing images can be viewed as a process of missing information reconstruction, which is a long-standing research problem. A comprehensive review of traditional cloud removal methods is provided in Shen et al. [6]. Based on the type of information it relies on, existing cloud removal methods can be roughly divided into four categories [7]: methods based on spatial information [8–12], methods based on multi-spectral information [13–16], methods based on multi-temporal information [17–19], and methods based on multi-source information [20,21].

Spatial-based methods regard cloud removal in remote sensing images as a straightforward image restoration task, and utilize information from cloud-free regions to restore corrupted pixels. They are mainly based on the following assumption: both cloud-covered and cloud-free regions share similar statistical and geometrical structures. The representative methods include interpolation methods [8], propagation-diffusion methods [9], variational-based methods [10], and example-based methods [11]. In recent years, many learning-based methods, especially Generative Adversarial Networks (GANs) [12] have been implemented in cloud removal tasks. While these direct restoration methods do not require additional data, they are typically only suitable for the restoration of small regions. When dealing with images containing extensive cloud-covered regions, spatial-based methods prove inadequate in restoring high-frequency texture details.

Multi-spectral-based methods recover the missing information primarily by using bands with higher penetration through the cloud. They reconstruct the destroyed bands mainly based on the assumption of a strong correlation between the spectral bands, e.g. using mathematical [13], physical [14], or GAN models [15,16]. However, they can only be applied in the case of haze and thin cirrus clouds. When all bands of an image are opaque, it is difficult to achieve high accuracy and robustness due to the lack of necessary reference data.

To address these issues, multi-temporal-based methods have emerged as effective solutions. These approaches aim to restore missing data by leveraging information from multiple time periods. They operate under the assumption that there are no major changes between data acquired at different time intervals. Therefore, corrupted data can be effectively recovered by utilizing cloud-free data from adjacent periods as a reference. These methods encompass replacement-based methods [17,22], filtering-based methods [18], and learning-based methods [19]. In contrast to others, these methods show remarkable superiority in cloud removal capability. However, these methods will not work in scenarios where significant changes have occurred in the ground objects within a short period of time, or where cloud-free reference data for adjacent time periods is not available (such as in regions with high cloud activity).

More recently, the multi-source approach has been developed, which uses images from additional sensors as auxiliary data for cloud removal purposes. Among these, SAR is particularly favoured due to its inherent advantages, including its ability to penetrate through clouds, and its all-weather and all-time capabilities [20]. As exemplified by recent works [21], SAR images can help restore texture details by compensating for the lack of information in cloud-covered areas of optical images.

However, certain limitations of SAR imagery should be recognised, including the presence of speckle noise, the lack of colour information, and the challenges associated with geometric distortion and shadows. These drawbacks make it difficult to distinguish different features in the same region. Moreover, the variance in imaging mechanism between SAR and optical modalities results in a considerable domain gap.

In response to the aforementioned challenges, we present a novel approach named the Multi-Level SAR-Guided Contextual Attention Network (MSGCA-Net). To address the problem of modality difference, the overall framework of MSGCA-Net is implemented as a parallel dual-stream network, which is used for optical and SAR image representation learning, respectively. Then, we design an SGCA module to enhance the information exchange between optical and SAR modalities. In this module, SAR image provides reliable global contextual information and authentic structure of cloud-covered regions, while the cloud-free regions of the optical image contribute to the reconstruction of texture details. In addition, we introduce a Multi-level Dynamical Mask (MDM) strategy to prevent the SGCA module from aggregating features in cloud-covered and cloud-shadow regions. Finally, the reconstructed features of the optical and SAR images are fused with the Gated Feature Fusion (GFF) module. Through qualitative and quantitative experimental evaluation on both simulated and real-world datasets, our proposed method demonstrates its effectiveness and superiority compared to several cloud removal methods.

In summary, our main contributions can be summarized as follows:

1. To leverage the dependable global contextual information inherent in SAR features, a SAR-Guided Contextual Attention (SGCA) module is designed. The module provides valuable guidance for capturing global interactions between contexts in order to maintain global consistency with the remaining cloud-free regions.
2. Furthermore, to prevent interference from cloud-covered and cloud-shadow regions during cloud removal estimation and to gradually restore large cloudy regions, we implement a strategy known as the Multi-level Dynamical Mask (MDM) strategy.
3. At last, we introduce the Gated Feature Fusion (GFF) module to transfer the complementary information, thereby achieving collaborative enhancement of both SAR and Optical features.

The paper is structured as follows: Section 2 provides a review of related works on cloud removal. In Section 3, our proposed method is described in detail. Section 4 presents the experimental setup and analyzes the results comprehensively. The effectiveness of each component in our method is discussed in Section 5. Finally, the conclusion of the paper is drawn in Section 6.

2. Related Works

2.1. SAR-Based Multi-Source Cloud Removal

In recent years, a number of studies [23,24] have delved into the potential of SAR-optical modality fusion for cloud removal purposes, which utilizes the complementary characteristics of the two modalities to estimate the cloudy region. Grohnfeldt et al. [25] proposed the first method using a cGAN network to directly fuse SAR and corrupted optical images to generate cloud-free results. Similarly, Meraner et al. [21] concatenated the SAR and optical image but used a deep residual neural network to estimate the cloud-free images. Gao et al. [26] and Darbagshahi et al. [27] employed a two-step approach with two GANs in series. In the initial step, the first GAN performed the task of translating SAR images into simulated optical counterparts. Subsequently, the second GAN integrated simulated optical images, original optical images, and SAR images to meticulously reconstruct the areas occluded by clouds.

However, current SAR-based cloud removal methods for multi-modal feature fusion lack intermodal information exchange. The SAR images can provide reliable global contextual information, while cloud-free regions in optical images can provide realistic texture details. Consequently, the effectiveness is constrained because the complementary information between modalities cannot be effectively transferred. To exploit the synergistic properties of dual imaging modalities, our approach incorporates the contribution of SAR to maintain global consistency and to guide the reconstruction of cloud-covered regions.

2.2. Attention-Based Method

Cloud removal can be viewed as a image restoration task, aiming to enhance the quality of images affected by clouds. Due to their superior performance, CNN-based methods were once the mainstay of image restoration. However, these methods generally suffer from a fundamental problem derived from their inherent local perceptual property, which makes it difficult to capture the long-range dependency. To make full use of contextual information, a series of methods based on contextual attention mechanism have been proposed. Yu et al. [28] proposed a contextual attention layer which replaces the features of missing regions with linear combinations of features from known regions. Zeng et al. [29] extended this idea by employing a pyramid of contextual attention at different hierarchical levels. In contrast to [29], Yi et al. [30] computed attention scores only once and reused them at multiple levels of abstraction, reducing the number of parameters and computational load.

The Transformer model, originally designed in [31], employs a self-attention mechanism to capture global interactions. This sophisticated framework has shown promising performance in image restoration tasks. Fang et al. [32] proposed a global-local fusion-based cloud removal (GLF-CR) algorithm that overcomes the limitations of local convolution. The method introduces the self-attention mechanism to establish global dependency and achieves a cloud removal effect beyond that of existing methods. Furthermore, Shuning et al. [33] proposed UFormer, which integrates a W-MSA module to improve the efficiency of transformer and reduce its computation cost.

In addition to transformers, recent advancements have introduced Mamba [34], a novel architecture inspired by classical state space models (SSM) [35], as an alternative approach with significant potential in handling long sequences efficiently. Unlike transformers, which suffer from quadratic complexity due to their self-attention mechanism, Mamba offers linear-time complexity and maintains global modeling capabilities, making it particularly suitable for processing large-scale remote sensing images. Recent works like Vim [36] and VMamba [37] have utilized State Space Models (SSM) to achieve linear complexity and enhance the effective receptive field for tasks such as image classification and segmentation on natural images. Vim performs selective scanning along the horizontal axis, while VMamba extends this by scanning in both horizontal and vertical directions, significantly impacting the receptive field in specific directions. However, these methods are less suited for very-high-resolution (VHR) remote sensing images, where spatial features can be multidirectional and more complex due to overhead perspectives. To address this challenge, RS-Mamba [38] leverages the linear complexity and global modeling ability of Mamba to handle entire images without cropping, preserving critical context information often lost in transformer-based approaches. By effectively modeling the intricate spatial relationships in VHR remote sensing data, RS-Mamba enhances performance in tasks such as object detection and semantic segmentation.

While Transformer-based models have achieved impressive results, their large number of parameters and extreme computational cost limit their in-depth exploration and practical application. Another issue is that the self-attention mechanism in transformers, while effective in high-level domains, is limited in low-level domains due to the constraints of fixed patch sizes. Inspired by the concept of contextual attention, we propose a SGCA module, which utilizes overlapable patches to improve feature aggregation, thus overcoming the limitations associated with fixed patch sizes.

3. Method

3.1. Overall Pipeline

As illustrated in Figure 1, the overall structure of the Multi-level SAR-Guided Contextual Attention Network (MSGCA-Net) is implemented as a concurrent two-stream network, which performs feature extraction and reconstruction for optical and SAR images, respectively. Each branch is a classical single-stage U-shaped architecture with skip connections and is finally fused at the fusion stage. The basic block we used is the NAF block, i.e., the Nonlinear Activation Free block, which is a simple but competitive baseline [39] for image restoration.

To take advantage of the reliable global contextual information of the SAR features, the optical features are reconstructed by the SAR-Guided Contextual Attention (SGCA) module. In addition, a dynamical mask strategy is introduced, which uses a masked-attention-matrix to prevent the SGCA module from aggregating features of cloud-covered regions. In order to achieve a gradual restoration of large cloudy regions, a multi-level feature reconstruction strategy of the optical features is executed within the decoding stage. We refer to the combination of these two strategies as the Multi-level Dynamical Mask (MDM) strategy. Finally, the reconstructed features of optical and SAR images are fused with the Gated Feature Fusion (GFF) module.

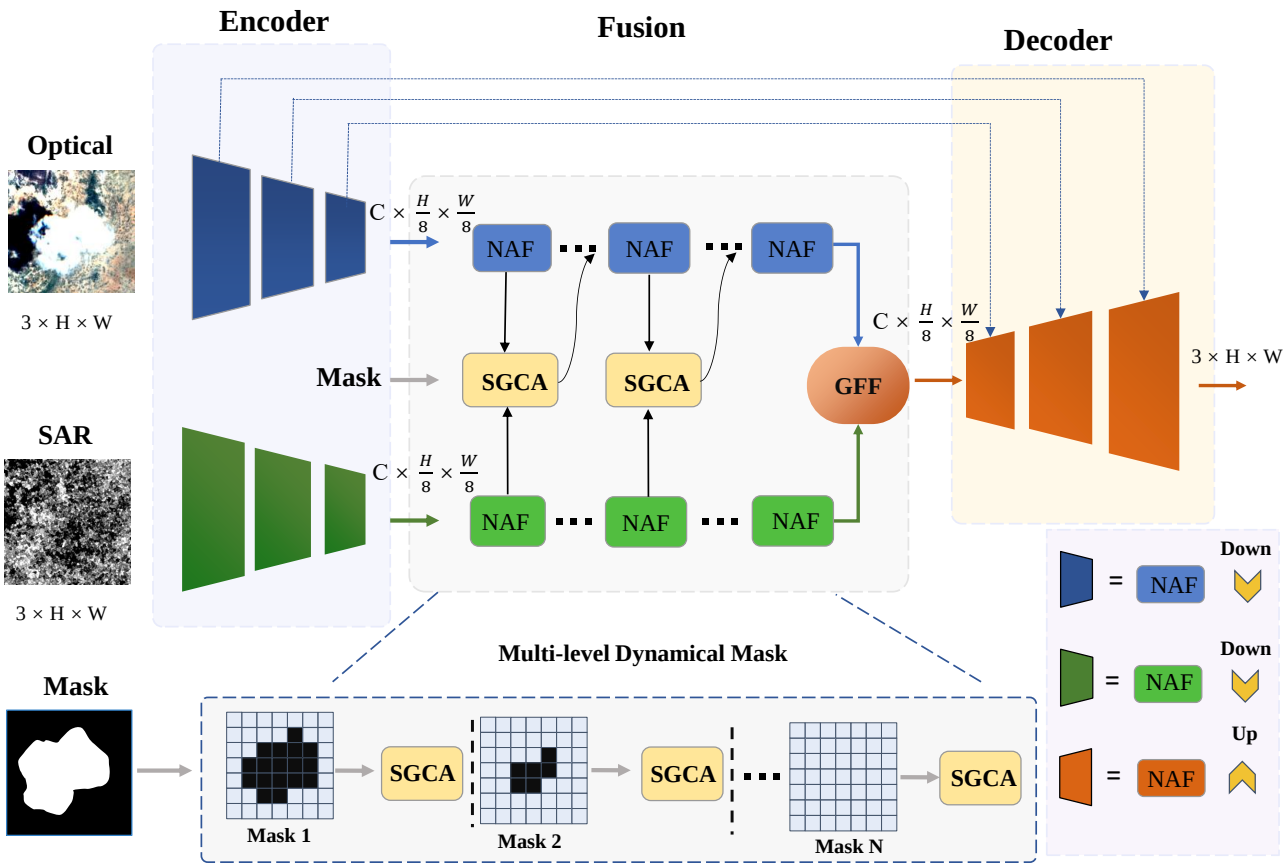

Figure 1. Overview of the proposed Multi-Level SAR-Guided Contextual Attention Network (MSGCA-Net) algorithm.

3.2. SAR Guided Contextual Attention

To better learn which cloud-free regions contribute to reconstructing cloudy regions, this module was designed. It is inspired by [28], which learns where to borrow or copy features from known background patches to generate missing patches in image restoration task. However, there is a problem with applying this method to the field of cloud removal: the absence of optical image information often leads to inaccurate similarity calculations between cloudy regions and cloud-free regions. This will result in inconsistent information inside and outside the restored region of a image.

Benefiting from SAR's immunity to cloud interference owing to its strong penetrative ability, this issue can be significantly alleviated. The features derived from SAR images retain reliable structural information, providing a reliable source of global contextual information. Consequently, they can offer valuable guidance for maintaining global consistency with other cloud-free regions.

Specifically, the feature map of optical branch is denoted as F_{opt} and the feature map of SAR branch is denoted as F_{sar}. As shown in Figure 2, we first partition the feature map F_{sar} into $K \times K$ patches with stride S and reshape them as convolutional filters F^i_{sar}. The stride S can be smaller than K, so it can obtain overlapping patches. Assuming that their number is N, the shape of F^i_{sar} is $\mathbb{R}^{K^2 \times C}$. To match cloudy patches F^i_{sar} with cloud-free ones F^j_{sar}, measurements can be performed with the normalized inner product (cosine similarity).

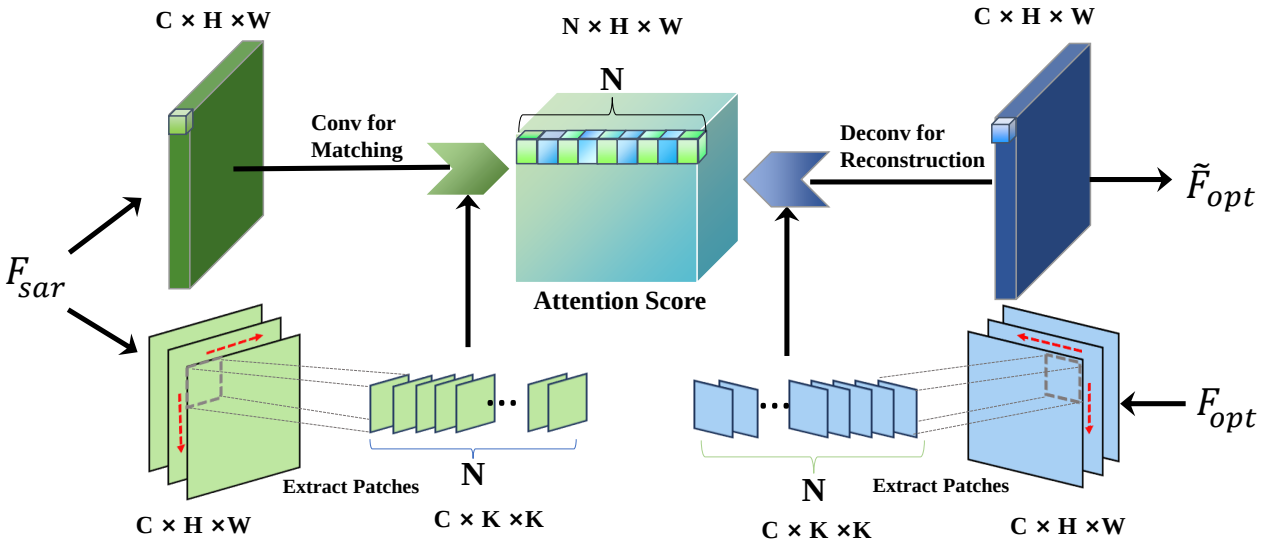

Figure 2. Overview of the SAR-Guided Contextual Attention (SGCA) module.

$$S_{i,j} = \langle \frac{F^i_{sar}}{\left\| F^i_{sar} \right\|}, \frac{F^j_{sar}}{\left\| F^j_{sar} \right\|} \rangle \tag{1}$$

where $S_{i,j}$ represents the similarity of cloudy patch F^i_{sar} and cloud-free patch F^j_{sar}. The attention matrix A is obtained by weighing the similarity using scaled softmax along the spatial dimension.

$$A_{i,j} = \frac{exp(S_{i,j})}{\sum_{j=1}^{N} exp(S_{i,j})} \tag{2}$$

where $A \epsilon \mathbb{R}^{N^2}$ is the attention matrix, and the N values in the i-th row represent the similarity between the i-th patch and all patches in the image.

Next, the feature map F_{opt} is partitioned into $(K \times K)$ patches in the same way as F_{sar}. Finally, the extracted patches F^j_{opt} are subjected to feature aggregation under the guidance of the attention matrix A, thus completing the reconstruction of the optical features:

$$\tilde{F}^i_{opt} = \sum_{j=1}^{N} F^j_{opt} \cdot A_{i,j} \tag{3}$$

where $\tilde{F}^i_{opt}$ is the i-th patch of the reconstructed features $\tilde{F}_{opt}$.

As demonstrated in Figure 2, the operations above are implemented as convolution, channel-wise softmax, and deconvolution, respectively.

3.3. Multi-Level Dynamical Mask

The SGCA module can effectively utilize the global contextual information from SAR images for cloud removal. However, we observe unstable optimisation when we use the above module for training. For large-scale clouds especially, it can even lead to gradient explosion. In addition, the SGCA module in the previous section relies only on the SAR features when computing the attention matrix, disregarding the position of the cloud. Therefore, the SGCA module may borrow invalid features from cloudy patches, which leads to undesirable results.

To handle the problem of image restoration where large holes are absent, Li et al. [40] introduced an innovative attention mechanism that is steered by an auxiliary mask. Similarly, we introduce the Masked-Attention-Matrix in Figure 3 to prevent the SGCA module from aggregating features of cloud-covered regions. With the cloud mask, the value in attention matrix A corresponding to the cloudy patches will be enforced as 0. The value of A can be refined under the guidance of cloud mask, which is formulated as:

$$A_{i,j} = \frac{exp(\lambda S_{i,j} + \tilde{M}_{i,j})}{\sum_{j=1}^{N} \lambda exp(S_{i,j} + \tilde{M}_{i,j})} \tag{4}$$

The mask $\tilde{M}_{i,j}$ is expressed as:

$$\tilde{M}_{i,j} \begin{cases} = 0, & \text{if patch p is cloud-free} \\ = -\tau, & \text{if patch p is cloudy} \end{cases} \tag{5}$$

where τ is a large positive integer (1000 in our experiments). Consequently, the values in attention matrix A corresponding to cloudy patches are nearly 0. Initially, the mask M is downsampled to match the size of the feature map.

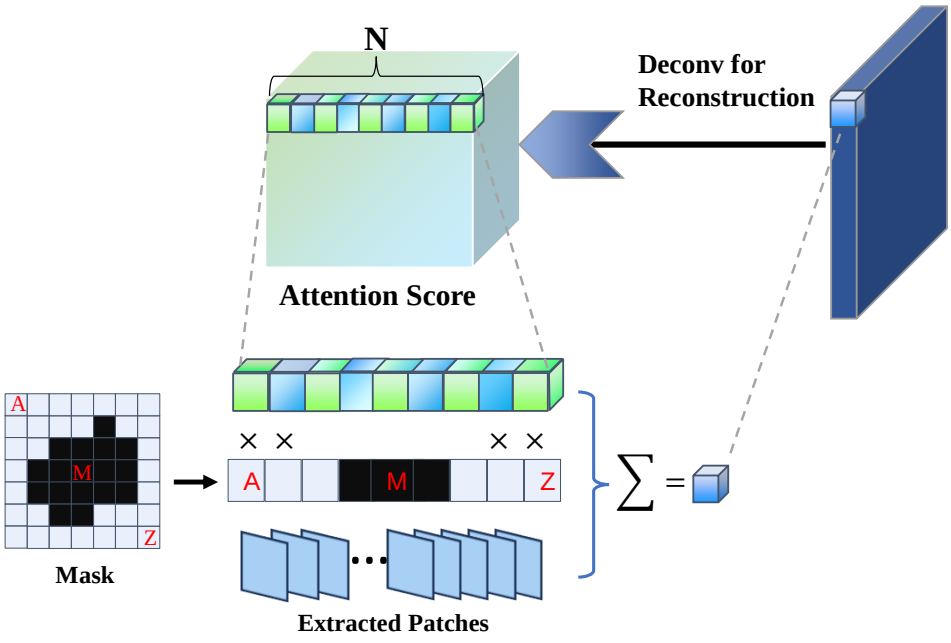

Figure 3. The Multi-level Dynamical Mask (MDM) strategy.

For cases with large cloudy regions, a multi-level feature reconstruction strategy is incorporated. The mask described above is initialised based on the input cloud mask and undergoes automatic updates throughout the feature reconstruction process. The updating follows an outside-in principle, i.e., a cloudy patch will be updated to a cloud-free patch in the next level as long as there is at least one cloud-free patch around it. If all patches around the patch are cloudy, they remain cloudy in the next level. As shown in the bottom of Figure 1, after multiple iterations of feature reconstruction, the mask is updated to become fully valid.

For images predominantly covered by clouds, the default attention strategy not only struggles to leverage the available information to reconstruct the cloudy regions, but also tends to underestimate the cloud-free pixels. Our "outside-in" principle allows for a gradual reconstruction of cloudy regions from the inside out, making the inpainting of large cloudy regions more feasible. The effectiveness of our design is manifested in Section 5.2.

3.4. Gated Feature Fusion

To better exploit the interactions among features of the optical and SAR image, a dual information propagation module is adopted. This module is derived from the module in [41], and has been adjusted to take into account the characteristics of cloud removal tasks. We add this module at the end of the Fusion stage, just before the Decoder stage commences. At this point, the optical features have been reconstructed in an ideal situation, thus enabling us to integrate the complementary advantages of the two modalities for feature integration. It exchanges messages between the features of the two modalities, in which a soft gating mechanism is exploited to control the rate. Figure 4 illustrates the GFF module.

Similarly, the feature map of the optical branch is denoted as F_{opt}, and the feature map of the SAR branch is denoted as F_{sar}. To enhance the optical feature by propagating complementary information from the SAR feature, a soft gating G_s, which controls the extent of information integration, is formulated as:

$$G_s = \sigma(Conv(Concat(F_{opt}, F_{sar}))) \tag{6}$$

where $Concat(\cdot)$ represents the operation of concatenation, $Conv(\cdot)$ denotes the mapping function executed by a convolution layer, and $\sigma(\cdot)$ represents the Sigmoid activation function. Utilizing G_s, we can adaptively merge F_{sar} into F_{opt}:

$$\tilde{F}_{opt} = \alpha(G_s \odot F_{sar}) \oplus F_{opt} \tag{7}$$

the operator $\odot$ represents the element-wise multiplication, whereas $\oplus$ stands for element-wise addition.

Symmetrically, the complementary information from SAR features is propagated to refine the optical feature as follows:

$$\begin{aligned} G_o &= \sigma(Conv(Concat(F_{opt}, F_{sar}))) \\ \tilde{F}_{sar} &= \beta(G_o \odot F_{opt}) \oplus F_{sar} \end{aligned} \tag{8}$$

Both α and β are trainable parameters that are initialized with a starting value of zero.

The GFF module requires processing both optical and SAR features concurrently, which doubles the computational complexity relative to processing a single feature branch. This increased complexity arises from the concatenation of F_{opt} and F_{sar}, the subsequent convolutional mapping, and the element-wise operations. However, since the feature fusion occurs only once throughout the network, its overall performance impact is negligible. This design ensures that the GFF module effectively balances computational cost with the adaptive integration of complementary information.

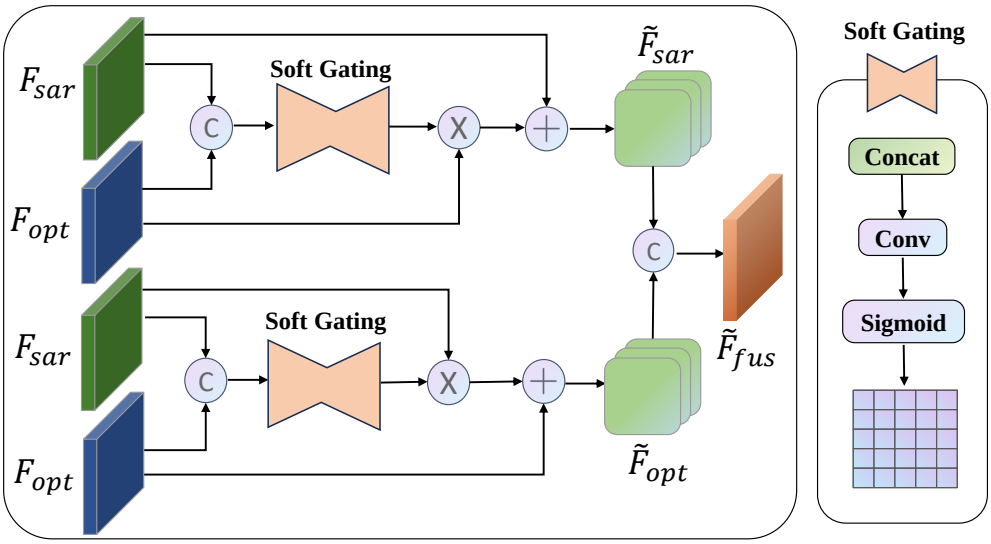

Figure 4. The Gated Feature Fusion (GFF) module, which is used to exchange messages between the two modalities. A soft gating is exploited to control the exchange rate, as shown on the right.

3.5. Loss Functions

As a classical loss function, $L1$ loss can quickly evaluate the quality of the generated image through intuitive pixel-by-pixel comparison. Thus, to effectively constrain the image quality, the L1 loss function is imported as follows:

$$L_1 = \|G(x) - y\|_1 \tag{9}$$

where y is the cloud-free image.

The structural similarity index measure (SSIM) index measures the similarity between two images and assesses quality based on the degradation of structural information. It is calculated from two images, x and y, as follows:

$$SSIM_{(x,y)} = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \cdot \frac{2\sigma_{xy} + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \tag{10}$$

where μ, σ^2, and σ_{xy} are the average, variance, and covariance, respectively. C_1 and C_2 are parameters for stabilizing the division with a weak denominator.

The $SSIM$ loss function in patch P can be expressed as

$$L_{SSIM} = \frac{1}{N}\sum(1 - SSIM_p) \tag{11}$$

where p is the center pixel of patch P. The size of the patch and Gaussian filter is 11×11.

Empirically, using only the L1 loss to evaluate the quality of generated images is too strict. Excessive emphasis on detailed information may result in neglecting the overall reconstruction quality of the image. Thus, we incorporate the $SSIM$ loss function alongside the $L1$ loss function to enhance the image quality. The final objective of our generators is computed using the following formula:

$$L_{all} = \lambda_1 L_1 + \lambda_2 L_{SSIM} \tag{12}$$

4. Experiments

4.1. Datasets and Metrics

In order to verify the feasibility of our method on the cloud removal task, both simulated and real-world experiments are conducted on different datasets. The detailed information of the dataset used in this paper is shown in Table 1. The results of cloud removal are evaluated with the normalized data based on peak signal-to-noise ratio (PSNR), SSIM, and mean absolute error (MAE).

The simulated experiments were conducted on the QXS-SAROPT dataset [42] (the QXS-SAROPT dataset under open access license CCBY is publicly available at https://github.com/yaoxu008/QXS-SAROPT, accessed on 21 April 2021.), which comprises 20,000 pairs of SAR-optical patches from diverse scenes in San Diego, Shanghai, and Qingdao. This dataset was constructed using freely available data acquired by the SAR satellite GaoFen-3 and optical satellites accessible via Google Earth. The SAR data utilized spotlight mode images with single polarization, while the optical images solely employed RGB bands. The size of each image patch was 256 × 256 pixels and the value of both the optical and SAR images was normalized to the range [0, 1].

Based on the QXS-SAROPT dataset, we simulated the clouds using a similar approach as Li et al. [7]. Specifically, images with a corrupted region smaller than 40% or larger than 30% were filtered out due to statistical evidence indicating that the proportion of land covered by clouds falls within this range.

Table 1. Detailed information of the datasets.

Dataset	Mode	Data Source	Band	Patch Size	Resolution Rg. × Az. (m)	Train/Val Number
SEN12MS-CR	Optical SAR	Sentinel-2 Sentinel-1	13 bands VV,VH	256	1 × 1	101,615 /8623
QXS-SAROPT	Optical SAR	Google Earth GaoFen-3	RGB single	256	1 × 1	16,000/4000
Flood-Zhengzhou	Optical SAR	Gaofen-1 Gaofen-3	RGB HH	512	1 × 1	1865/467

Real-world experiments were conducted employing the large-scale dataset SEN12MS-CR. The SEN12MS-CR dataset comprises Sentinel satellite images from the Copernicus project, providing comprehensive coverage of geographical and meteorological conditions across all continents, as well as various seasons. The dataset consists of an dual-polarized (VV and VH) SAR image from Sentinel-1, a cloud-free multi-spectral image from Sentinel-2, and a cloudy multi-spectral image also from Sentinel-2. All images are ortho-corrected and geographically registered. Notably, the acquisition times for the cloudless and cloudy images are closely aligned. The VV and VH polarization images of the SAR image are normalized to the range [−25, 0] and [−32.5, 0], respectively, and then rescaled to a range of [0, 1]. Similarly, all bands of the optical images are clipped to values within [0, 10,000], and subsequently rescaled to a range of [0, 1].

To further validate the cloud removal capability of our model in real-world scenarios, we created a specialized dataset named Flood-Zhengzhou. The dataset covers approximately 2445 square kilometers of Zhengzhou, Henan, China. The dataset underwent meticulous preprocessing to ensure its quality and relevance. This preprocessing included essential steps such as ortho-correction and geographic registration. The data were sliced according to the size of 512× 512 and a total of 2332 triple samples are generated. Each sample included a Gaofen-3 dual-polarized HH SAR image, a Gaofen-1 cloudy RGB image captured on 24 July 2021 and a Gaofen-1 cloud-free RGB image captured on 6 December 2021. Notably, all pixel values across the images were normalized to the range [0, 1].

4.2. Implementation Details

The proposed method was implemented using publicly available Pytorch and trained end-to-end on four NVIDIA GeForce RTX 3090 GPUs. The hyperparameters of training were set as follows: learning rate $\eta = 1 \times 10^4, \lambda_1 = \lambda_2 = 1$. After the experiment, the batch size was set to 8 and the maximum epoch of training was set to 100. The optimizer we used was AdamW, with the following momentum parameters: $\beta 1 = 0.9$, $\beta 2 = 0.9, weight\_decay = 0$. The initial learning rate was gradually reduced from $1e^{-4}$ to $1e^{-7}$ with the cosine annealing schedule. In the training stage, the samples were randomly cropped into 192 × 192 patches. Balancing the model's performance against its complexity, the number of the NAF block was set to 36. The number of the SGCA module, i.e., the number of mask updates, was set to 3 (the codes of our model will be released at: https://github.com/javey-q/MSGCA-Net, accessed on 13 December 2024).

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

The proposed MSGCA-Net method was compared with the state-of-the-art cloud removal methods, including SpA GAN (2020) [43], SAR-Opt-cGAN (2018) [25], DSen2-CR (2020) [21], DRIB GAN (2022) [27], GLF-CR (2022) [32], and Align-CR (2023) [44].

SpA GAN takes all channels of optical images as inputs and uses a spatial attention network (SPANet) as a generator to achieve the transformation from cloudy images to cloud-free images. Both SAR-Opt-cGAN and DSen2-CR use SAR images as priors to assist in the reconstruction process under optically impervious clouds. SAR-Opt-cGAN is extended from U-Net, while DSen2-CR is extended from the residual dense network [45]. In these two methods, the SAR band and the optical band are fused by simple concatenation to estimate all the bands of the optical image. DRIB-GAN first uses the first GAN network to translate SAR images into optical images, resulting in preliminary estimation images. Then, the estimation images, SAR images, and cloud-covered optical images are concatenated together and input into the second GAN network to obtain the final reconstruction result. Align-CR design a model for cloud removal in high-resolution optical images guided by low-resolution SAR images, using implicit alignment of multi-modal and multi-resolution data to enhance performance. In particular, GLF-CR designs wo parallel branches for optical and SAR image feature representation, respectively. Considering the important role of global information, it proposes a novel global-local fusion based algorithm to exploit the complementary information in SAR images.

In Table 2, the overall comparison results of the MSGCA-Net with the above-mentioned methods on the simulated dataset (QXS-SAROPT) and the real-world dataset (SEN12MS-CR) are given. The above-mentioned GLF-CR represents the SOTA of the SAR-based cloud removal method. We marked the highest score in each index in bold. The proposed MSGCA-Net achieved 29.2 dB PSNR and 0.905 SSIM index on the SEN12MS-CR dataset, exceeding the previous SOTA 0.2 dB and 0.02 points respectively. On the QXS-SAROPT dataset, the PSNR metric of MSGCA-Net was 0.91 dB higher, and the SSIM metric was 0.011 points higher than the second place GLF-CR. DSen2-CR and DRIB-GAN ranked significantly lower than GLF-CR and our method in PSNR metric, because they simply concatenated the optical and SAR images, ignoring the large domain differences between the multimodal data. The quantitative results demonstrate that our method outperforms other comparative methods significantly on both datasets.

The visualization comparison of the cloud removal performance of each method is shown in Figure 5. From left to right, each column represents the cloudy image, the cloud removal result from SpA GAN, SAR-Opt-cGAN, DSen2-CR, DRIB GAN, GLF-CR, our method, and the cloud-free image.

Table 2. Quantitative comparisons of proposed method to state-of-the-art methods.

Method	Params	MAC	QXS-SAROPT			SEN12MS-CR		
			PSNR	SSIM	MAE	PSNR	SSIM	MAE
SpA GAN (2020) [43]	0.42	33.96	21.28	0.824	0.0355	24.86	0.753	0.0444
SAR-Opt-cGAN (2018) [25]	170.0	26.12	22.87	0.827	0.0323	25.29	0.759	0.0441
DSen2-CR (2020) [21]	18.9	1238.0	24.34	0.882	0.0245	27.37	0.870	0.0319
DRIB GAN (2022) [27]	37.05	42.69	24.51	0.878	0.0286	-	-	-
GLF-CR (2022) [32]	14.65	243.0	27.04	0.894	0.0171	29.07	0.885	0.0266
Align-CR (2023) [44]	43.51	447.7	26.95	0.898	0.0161	-	-	-
MSGCA-Net (Ours)	63.3	121.4	**27.95**	**0.905**	**0.0153**	**29.2**	**0.905**	**0.0215**

Among these methods, SpA GAN achieved the worst results, due to the fact that it only utilizes information from optical images. This indicates that the ultilization of SAR images is of vital importance in cloud removal. SAR-Opt-cGAN achieved slightly better results than SpA GAN, but also generated undesirable contents, especially for cloud-covered regions. DSen2-CR and DRIB-GAN both leverage the SAR image as a form of prior to guide the reconstruction process, so they achieved relatively reasonable reconstruction content. However, both methods involve the direct concatenation of SAR with optical images and the extraction of features without taking into account the inter-modal domain differences, resulting in a limited gain being achieved. Unlike DSen2-CR and DRIB-GAN, GLF-CR incorporates global contextual interactions that take into account the information embedded in cloud-free regions. It can maintain global consistency between the cloud removal result and the surrounding cloud-free regions. However, as shown in the red box area, the reconstructed image of this method still exhibits blurriness compared to our method. This illustrates the capability of our method to effectively leverage the global contextual information embedded in SAR features and subsequently steer the interactions of global optical features.

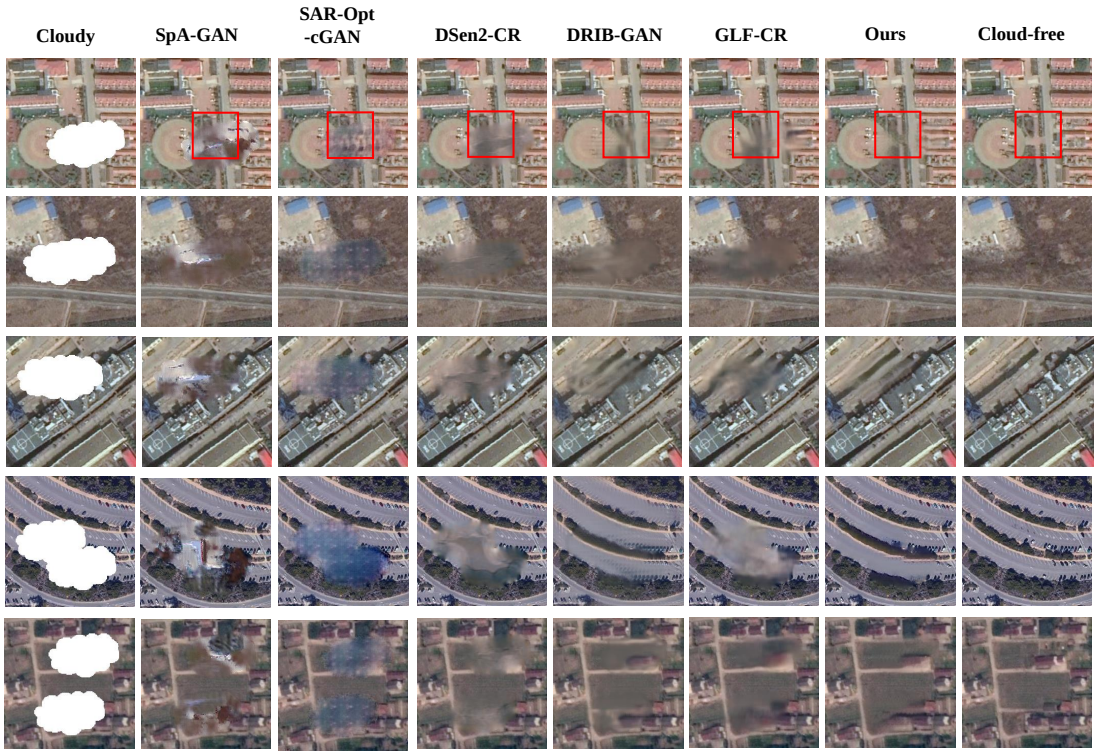

Figure 5. Qualitative comparisons of proposed method to state-of-the-art methods. The red box in the figure highlights a specific region or detail that is of particular importance.

To further demonstrate the performance of our proposed method, we also present its results in real-world scenarios. In Figure 6, the restoration results of RGB images from the SEN12MS-CR dataset are shown. From left to right, each column represents the optical cloudy image, the SAR image, the cloud removal result of our method, and the reference optical image at other times. It can be observed that our algorithm achieves precise restoration when dealing with thin clouds or areas with less cloud coverage. However, as cloud coverage increases, some regions in our recovery results start to become slightly blurred.

In Figure 7, the restoration results of RGB images in Flood-Zhengzhou dataset are provided. From left to right, each column represents the optical cloudy image, the SAR image, the cloud removal result of our method and the reference optical image of other time. By comparing our cloud removal result with the reference image, it is evident that although our method cannot perfectly restore the texture details, it can accurately reconstruct the structures of the missing region.

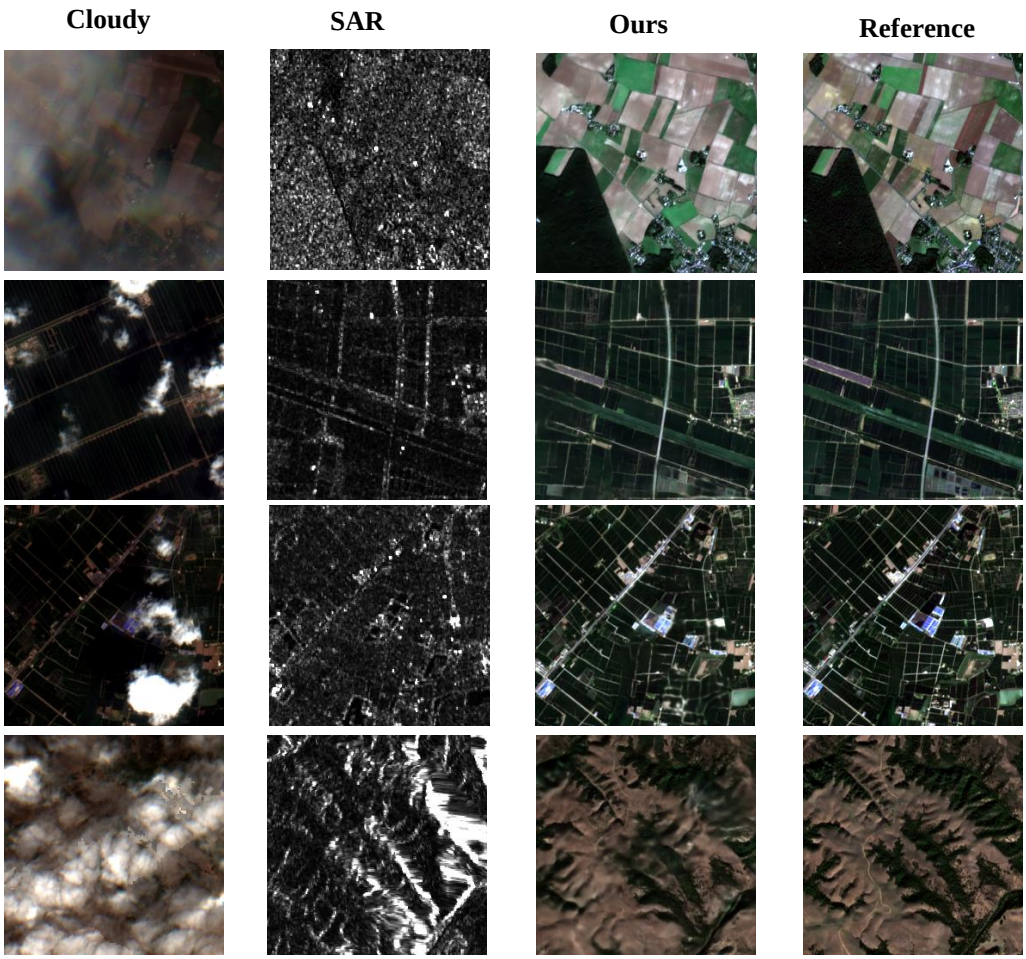

Figure 6. Cloud removal results of our method in SEN12MS-CR dataset. From **left** to **right**, each column represents the cloudy image, the SAR image, the cloud removal result, and the cloud-free reference image.

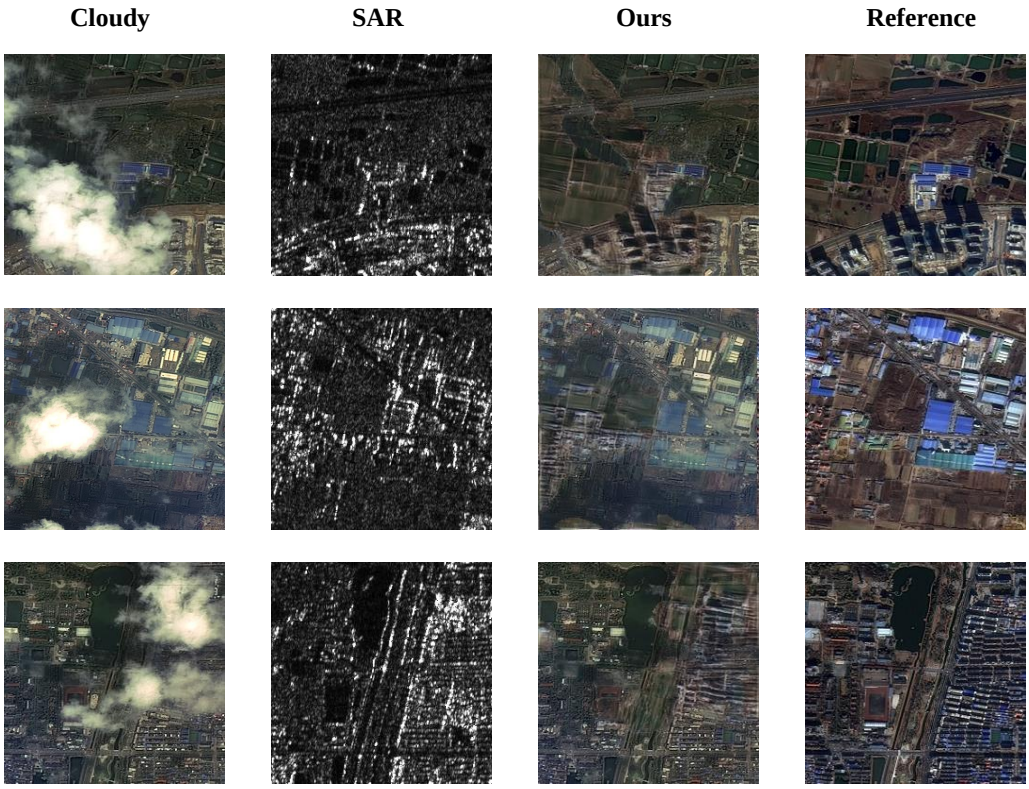

Figure 7. Cloud removal results of our method in Flood-Zhengzhou dataset. From **left** to **right**, each column represents the cloudy image, the SAR image, the cloud removal result, and the cloud-free reference image.

5. Discussion

5.1. Analysis of Computational Cost

In the practical application of remote sensing, the computational complexity of a model is also an important reference index. Therefore, we compared the model's parameters (Params) and Memory Access Cost (MAC), which represents the total amount of memory exchanged when the model completes one forward pass for a single sample (one image). It quantifies the spatial complexity of the model, measured in bytes. The MACs are estimated by an input with a spatial size of 256 × 256. The quantitative comparisons are shown in Table 2. The overall comparison of PSNR (SSIM) vs. computational cost is visualized in Figure 8.

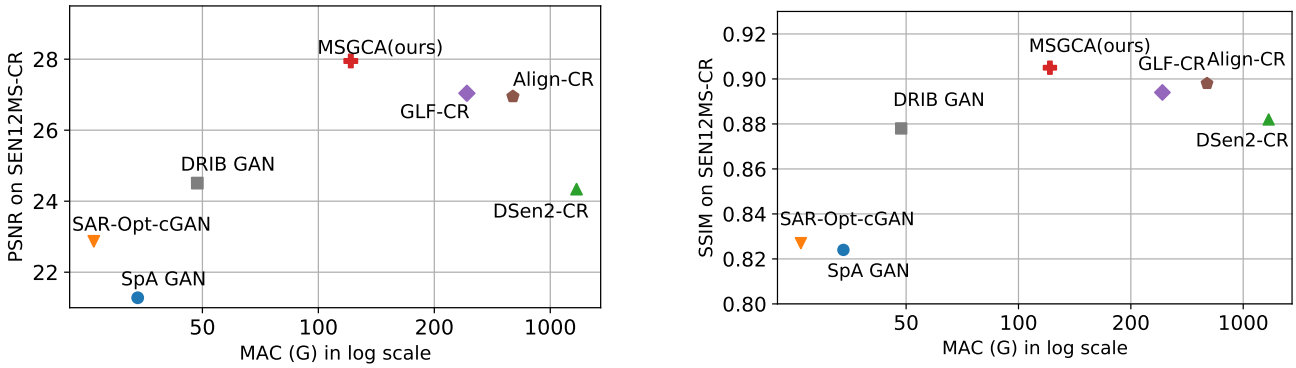

Figure 8. PSNR vs. computational cost (**left**) and SSIM vs. computational cost (**right**) on QXS-SAROPT Dataset.

In Figure 8, the vertical axis represents the PSNR/SSIM index, which quantifies the image quality, while the the horizontal axis denotes the MACs, which indicates computational cost. The exceptional performance of a model is indicated by its proximity to the top-left corner of the plot. As can be seen in the figure, the proposed method shows comparable performance to GLF-CR in terms of PSNR and SSIM metrics, while having a lower MAC. This suggests that the proposed method achieves SOTA in image quality metrics while maintaining low computational complexity.

5.2. Ablation Study

The proposed method includes three main operations: SGCA, MDM, and GFF. To further analyze the role of each operation, we performed ablation experiments sequentially. The quantitative and qualitative results on the QXS-SAROPT dataset are shown in Table 3 and Figure 9, where "w/o" is an abbreviation for "without".

Table 3. Quantitative comparisons of different variants of our method.

Method	PSNR	SSIM	MAE
w/o SGCA	25.96	0.880	0.0178
w/o MDM	27.11	0.890	0.0168
w/o GFF	27.42	0.898	0.0166
Ours	27.95	0.905	0.0163

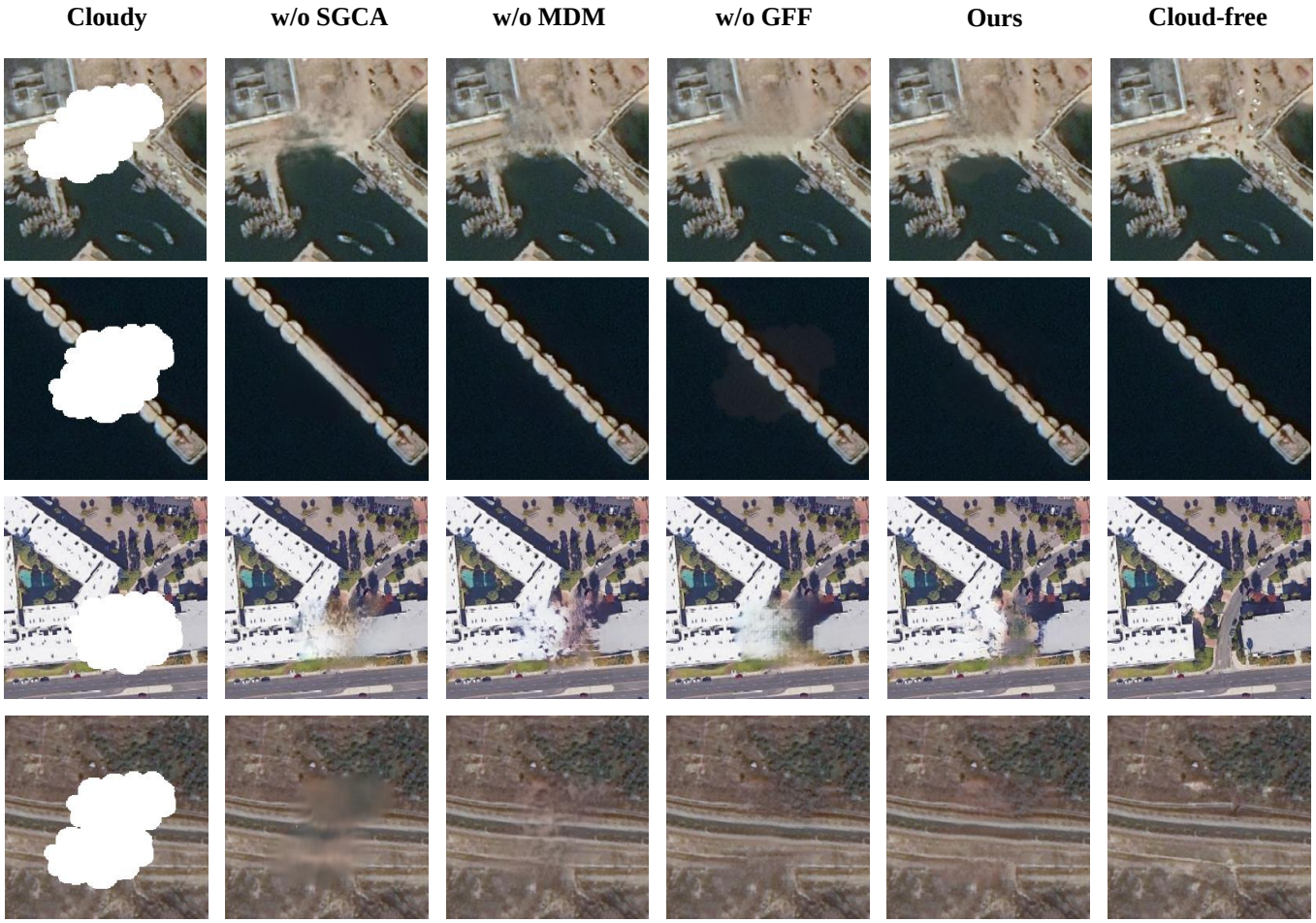

Figure 9. Qualitative comparisons of different variants of our method.

Effects of the SAR-Guided Contextual Attention: To verify the effects of SGCA module, we simply removed the module from the network, denoted as 'w/o SGCA'. Since SAR images are not affected by cloud cover, they can provide reliable global contextual information. This point can be validated by comparing the result of 'w/o SGCA' and 'Ours'. As shown in the second row in Figure 9, the model without the SGCA module was unable to effectively reconstruct regions with repetitive structures due to the absence of global information provided by SAR. It can be seen that the SGCA module provides valuable guidance, capturing global interactions between contexts and making the estimated image more consistent with ground truth. The quantitative results in Table 3 also validate its necessity.

Effects of the Multi-level Dynamical Mask strategy: To verify the effects of the MDM strategy, we conducted experiments by removing this strategy from the network, denoted as 'w/o MDM'. As observed from the fourth scene in Figure 9, when cloud regions become extensive, the 'w/o MDM' approach tended to erroneously utilize features from the cloudy region for the reconstruction, resulting in an unreasonable restoration result. Specifically, it failed to accurately recover structural information such as road details within the cloud-covered areas. In contrast, by incorporating the MDM strategy, our model can restore images with more detail and fewer artifacts, ensuring a more reasonable and accurate restoration of large cloud-covered regions.

Effects of the Gated Feature Fusion: To validate the effects of the GFF, we substituted the GFF with a simple concatenation operation followed by a normal convolution layer, denoted as 'w/o GFF'. Figure 9 visually demonstrates that the proposed network excelled in recovering more complete texture structures, yielding superior visual effects. For a more detailed and explicit comparative analysis, we include in Table 3 the quantitative results. These results indicate that the inclusion of the GFF module contributes significantly to enhancing the performance of the evaluation metrics. The comparison experiment showcases that incorporating the GFF into the model enables efficient utilization of multi-modal data fusion.

5.3. Limitations and Shortcomings

While our proposed method demonstrates significant effectiveness in various scenarios, it also has certain limitations and shortcomings that warrant further exploration.

Fixed Number of Mask Updates: The iterative updating of the mask is designed to progressively remove clouds from the inside out, ensuring a more natural restoration of obscured areas. We chose a fixed number of updates (set to 3) rather than continuing updates until there are no zero values in the matrix because a dynamic stopping criterion could introduce unpredictability into the training process, potentially complicating convergence and stability. While this approach has proven effective under conditions where cloud coverage ranges from 30% to 40%, setting the number of mask updates to a fixed value may present limitations, especially in scenarios with extensive cloud coverage. In extreme situations involving large cloudy regions, three iterations might not be sufficient, which could limit the effectiveness of the de-clouding process. Future work could explore adaptive mechanisms that dynamically adjust the number of iterations based on the extent of cloud coverage detected within the image, thereby enhancing the robustness of the method.

Over-reliance on SAR Image Features: The SGCA module exhibits an over-reliance on SAR image features. For regions where SAR images cannot provide effective information, such as densely vegetated mountainous areas, this module has limitations. In these environments, SAR images typically exhibit low backscatter coefficients, appearing as dark areas that are less suitable for reconstructing global features. This characteristic poses challenges for the method, particularly in scenarios with extensive cloud coverage over dense vegetation. The lack of reliable global information from SAR images in these environments can lead to suboptimal results. Future research will focus on mitigating this dependency by exploring methods that can better leverage available structural information from optical images alone or integrating auxiliary data sources like historical imagery and digital elevation models to enhance performance in challenging scenarios.

6. Conclusions

In this article, we propose a novel approach named the Multi-Level SAR-Guided Contextual Attention Network (MSGCA-Net) to enhance the fusion of SAR and optical features. In the framework of MSGCA-Net, the SAR image provides reliable global contextual information and the genuine structure of cloud-covered regions, while the cloud-free regions of the optical image contribute to the reconstruction of texture details. We evaluated the performance of our model both on simulated and real-world datasets and then compared their competition cost. Compared with other cloud removal methods, our method performs better in both qualitative and quantitative evaluation, which proves that our method is effective and superior.

There was an observed gap in the performance of our model on real datasets compared to the results on the simulated datasets because the simulated datasets can guarantee a strict ground truth by superimposing synthetic cloud cover on cloud-free images. However, models trained by simulation clouds are difficult to apply in real scenarios. Providing a solution to this issue can be discussed in future work.

Author Contributions: Conceptualization, G.L. and J.Q.; methodology, G.L. and J.Q.; validation, G.L. and J.Q.; formal analysis, G.L. and J.Q.; investigation, G.L. and J.Q.; writing—original draft preparation, G.L. and J.Q.; writing—review and editing, G.L., J.Q. and Y.Y; supervision, G.L.and Y.Y.; All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Key Research and Development Program of China under Grant 2022ZD0160401 and in part by the National Natural Science Foundation of China under Grant 62273282.

Data Availability Statement: The SEN12MS-CR dataset used in this study is available at https://mediatum.ub.tum.de/1554803, accessed on 2 July 2020, and the QXS-SAROPT dataset is available at https://github.com/yaoxu008/QXS-SAROPT, accessed on 21 April 2021.

Acknowledgments: The authors would like to thank the editors and anonymous reviewers for their valuable comments and suggestions, as well as those researchers who make public codes and public datasets.

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

References

1. Lin, Z.; Leng, B. SSN: Scale Selection Network for Multi-Scale Object Detection in Remote Sensing Images. *Remote Sens.* **2024**, *16*, 3697. [CrossRef]
2. Mo, N.; Zhu, R. Semi-Supervised Subcategory Centroid Alignment-Based Scene Classification for High-Resolution Remote Sensing Images. *Remote Sens.* **2024**, *16*, 3728. [CrossRef]
3. Liu, G.; Li, C.; Zhang, S.; Yuan, Y. VL-MFL: UAV Visual Localization Based on Multi-Source Image Feature Learning. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62* , 5618612. [CrossRef]
4. Shen, J.; Huo, C.; Xiang, S. Siamese InternImage for Change Detection. *Remote Sens.* **2024**, *16*, 3642. [CrossRef]
5. King, M.D.; Platnick, S.; Menzel, W.P.; Ackerman, S.A.; Hubanks, P.A. Spatial and temporal distribution of clouds observed by MODIS onboard the Terra and Aqua satellites. *IEEE Trans. Geosci. Remote Sens.* **2013**, *51*, 3826–3852. [CrossRef]
6. 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]
7. Li, W.; Li, Y.; Chan, J.C.W. Thick Cloud Removal with Optical and SAR Imagery via Convolutional-Mapping-Deconvolutional Network. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 2865–2879. [CrossRef]
8. Zhang, C.; Li, W.; Travis, D. Gaps-fill of SLC-off Landsat ETM+ satellite image using a geostatistical approach. *Int. J. Remote Sens.* **2007**, *28*, 5103–5122. [CrossRef]
9. 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]
10. 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]
11. Criminisi, A.; Pérez, P.; Toyama, K. Region filling and object removal by exemplar-based image inpainting. *IEEE Trans. Image Process.* **2004**, *13*, 1200–1212. [CrossRef]
12. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 2672–2680.

13. 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]
14. Lv, H.; Wang, Y.; Shen, Y. An empirical and radiative transfer model based algorithm to remove thin clouds in visible bands. *Remote Sens. Environ.* **2016**, *179*, 183–195. [CrossRef]
15. 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, 21–26 July 2017; pp. 48–56.
16. 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]
17. Lin, C.H.; Tsai, P.H.; Lai, K.H.; Chen, J.Y. Cloud removal from multitemporal satellite images using information cloning. *IEEE Trans. Geosci. Remote Sens.* **2012**, *51*, 232–241. [CrossRef]
18. Chen, J.; Jönsson, P.; Tamura, M.; Gu, Z.; Matsushita, B.; Eklundh, L. A simple method for reconstructing a high-quality NDVI time-series data set based on the Savitzky–Golay filter. *Remote Sens. Environ.* **2004**, *91*, 332–344. [CrossRef]
19. Li, X.; Shen, H.; Zhang, L.; Zhang, H.; Yuan, Q.; Yang, G. Recovering quantitative remote sensing products contaminated by thick clouds and shadows using multitemporal dictionary learning. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 7086–7098.
20. Bamler, R. Principles of synthetic aperture radar. *Surv. Geophys.* **2000**, *21*, 147–157. [CrossRef]
21. 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]
22. Zeng, C.; Shen, H.; Zhang, L. Recovering missing pixels for Landsat ETM+ SLC-off imagery using multi-temporal regression analysis and a regularization method. *Remote Sens. Environ.* **2013**, *131*, 182–194. [CrossRef]
23. 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*, 1–15. [CrossRef]
24. 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*, 1–14. [CrossRef]
25. 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, IEEE, Valencia, Spain, 22–27 July 2018; pp. 1726–1729.
26. 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]
27. 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.* **2022**, *60*, 1–9. [CrossRef]
28. Yu, J.; Lin, Z.; Yang, J.; Shen, X.; Lu, X.; Huang, T.S. Generative Image Inpainting with Contextual Attention. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5505–5514.
29. Zeng, Y.; Lin, Z.; Lu, H.; Patel, V.M. CR-Fill: Generative Image Inpainting with Auxiliary Contextual Reconstruction. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 14164–14173.
30. Yi, Z.; Tang, Q.; Azizi, S.; Jang, D.; Xu, Z. Contextual Residual Aggregation for Ultra High-Resolution Image Inpainting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 7508–7517.
31. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30* , 5998–6008.
32. 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]
33. Han, S.; Wang, J.; Zhang, S. Former-CR: A Transformer-Based Thick Cloud Removal Method with Optical and SAR Imagery. *Remote Sens.* **2023**, *15*, 1196. [CrossRef]
34. Gu, A.; Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv* **2023**, arXiv:2312.00752.
35. Gu, A.; Goel, K.; Ré, C. Efficiently modeling long sequences with structured state spaces. *arXiv* **2021**, arXiv:2111.00396.
36. Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; Wang, X. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv* **2024**, arXiv:2401.09417.
37. Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Liu, Y. VMamba: Visual State Space Model. *arXiv* **2024**, arXiv:2401.10166.
38. Zhao, S.; Chen, H.; Zhang, X.; Xiao, P.; Bai, L.; Ouyang, W. Rs-mamba for large remote sensing image dense prediction. *arXiv* **2024**, arXiv:2404.02668. [CrossRef]
39. Chen, L.; Chu, X.; Zhang, X.; Sun, J. Simple Baselines for Image Restoration. In *Proceedings of the Computer Vision–ECCV, Tel Aviv, Israel, 23–27 October 2022*; Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T., Eds.; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2022; pp. 17–33. [CrossRef]
40. Li, W.; Lin, Z.; Zhou, K.; Qi, L.; Wang, Y.; Jia, J. MAT: Mask-Aware Transformer for Large Hole Image Inpainting. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 10748–10758. [CrossRef]

41. Guo, X.; Yang, H.; Huang, D. Image Inpainting via Conditional Texture and Structure Dual Generation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 14134–14143.
42. Huang, M.; Xu, Y.; Qian, L.; Shi, W.; Zhang, Y.; Bao, W.; Wang, N.; Liu, X.; Xiang, X. The QXS-SAROPT Dataset for Deep Learning in SAR-Optical Data Fusion. *arXiv* **2021**, arXiv:2103.08259.
43. Pan, H. Cloud Removal for Remote Sensing Imagery via Spatial Attention Generative Adversarial Network. *arXiv* **2020**, arXiv:2009.13015.
44. Xu, F.; Shi, Y.; Ebel, P.; Yang, W.; Zhu, X.X. High-resolution cloud removal with multi-modal and multi-resolution data fusion: A new baseline and benchmark. *arXiv* **2023**, arXiv:2301.03432.
45. Zhang, Y.; Tian, Y.; Kong, Y.; Zhong, B.; Fu, Y. Residual Dense Network for Image Super-Resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2472–2481.