remote sensing

MDPI

Article

STAIT: A Spatio-Temporal Alternating Iterative Transformer for Multi-Temporal Remote Sensing Image Cloud Removal

Yukun Cui [1], Jiangshe Zhang [1,*], Haowen Bai [1], Zixiang Zhao [1], Lilun Deng [1], Shuang Xu [2] and Chunxia Zhang [1]

[1] School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an 710049, China; cuiyukun@stu.xjtu.edu.cn (Y.C.); hwbaii@stu.xjtu.edu.cn (H.B.); zixiangzhao@stu.xjtu.edu.cn (Z.Z.); statdll@stu.xjtu.edu.cn (L.D.); cxzhang@mail.xjtu.edu.cn (C.Z.)
[2] School of Mathematics and Statistics, Northwestern Polytechnical University, Xi'an 710072, China; xs@nwpu.edu.cn
* Correspondence: jszhang@mail.xjtu.edu.cn

Highlights

What are the main findings?

- A novel Spatio-Temporal Alternating Iterative Transformer (STAIT) is proposed to explicitly model the dynamic dependencies in multi-temporal remote sensing image cloud removal task.
- An efficient framework combining multi-level feature extraction and a weight-sharing decoder is designed to ensure high-quality, temporally consistent reconstruction.

What are the implication of the main findings?

- The method significantly improves cloud removal accuracy, effectively restoring surface details obscured by thick clouds.
- It provides a robust and efficient solution for generating continuous remote sensing data, enhancing the reliability of Earth observation applications.

Abstract

Multi-temporal remote sensing image cloud removal aims to reconstruct land surface information in regions obscured by clouds and their shadows, thereby mitigating a major constraint on the application of remote sensing imagery. However, existing multi-temporal deep learning methods for cloud removal often fail to model complex spatio-temporal dynamics, leading to suboptimal performance. To address this challenge, we propose a novel framework for multi-temporal cloud removal. In this architecture, the most critical component is the Spatio-Temporal Alternating Iterative Transformer (STAIT), which primarily consists of temporal and spatial attention mechanisms. STAIT is engineered to refine spatio-temporal feature representation by establishing an effective interplay between spatial details and temporal dynamics. Our framework is enhanced by an efficient image token generator with group convolution-based multi-level feature extraction to manage complexity, and a pixel reconstruction decoder with a shared progressive upsampling network to improve reconstruction by learning time-invariant features. Experimental results demonstrate that by explicitly modeling spatio-temporal feature dependencies, our approach achieves superior performance in restoring high-fidelity, cloud-free imagery.

Keywords: remote sensing image; multi-temporal cloud removal; transformer

Check for updates

Academic Editor: Haopeng Zhang

Received: 29 December 2025
Revised: 10 February 2026
Accepted: 11 February 2026
Published: 14 February 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.

1. Introduction

Optical satellite remote sensing serves as a fundamental technology for observing and comprehending our planet. It enables a wide range of applications, including precision agriculture [1,2], urban dynamic monitoring [3,4], and environmental change tracking [5,6]. However, the pervasive presence of clouds poses severe obstacles to fundamental remote sensing tasks like ground target recognition [7] and image retrieval [8,9]. Furthermore, it plays a crucial role in low-level multi-spectral image reconstruction [10–13], providing high-quality foundational data for various downstream quantitative analyses. This widespread "cloud interference" phenomenon not only significantly impedes research and decision-making efficiency in critical application domains, but also leads to incomplete datasets and discontinuous time series, thereby compromising the validity of subsequent analyses and potentially leading to flawed interpretations.

The widespread presence of clouds makes simple single-shot data collection ineffective. Therefore, high-fidelity remote sensing usually requires more complex and advanced imaging methods. A prevalent methodology relies on the temporal correlation between multi-temporal images to reconstruct the missing land surface data under thick clouds [14–16]. This approach generally relies on multi-temporal observations from a single platform for data acquisition. By inputting a sequence of partially cloudy remote sensing images acquired at short time intervals, it aims to predict cloud-free remote sensing imagery. In practical applications, this methodology requires the accumulation and processing of substantial historical image data to fully exploit temporal variations and consistencies for cloud removal.

Current multi-temporal cloud removal techniques are primarily divided into tensor completion, generative models, and feature-learning frameworks. However, tensor-based optimization often lacks flexibility in handling complex occlusions, while generative approaches (e.g., GANs, DDPMs) face challenges regarding training stability and inference efficiency. Although deep feature-learning methods leveraging CNNs and Transformers have shown promise in capturing contextual dependencies, they share a critical limitation: most approaches simply concatenate images from multiple time nodes as inputs to attention mechanisms. Consequently, they fail to explicitly model the underlying spatio-temporal feature dependencies and the inherent evolutionary relationships across temporal sequences. In other words, while these methods leverage multi-temporal information, they lack a deep, explicit spatio-temporal modeling of the potential dependencies and correlations between cloudy images. Such a simplistic approach of channel-wise concatenation essentially treats temporal snapshots as isolated inputs, neglecting the inherent spatio-temporal continuity and the complex dependencies that connect them. Consequently, the model struggles to capture crucial spatio-temporal dynamics and feature dependencies. Furthermore, the absence of explicit constraints on these dependencies can result in temporally inconsistent predictions, which in turn hampers the overall predictive performance and generalization ability.

To address this deficiency, we propose a novel Transformer module, Spatio-Temporal Alternating Iterative Transformer (STAIT), to explicitly model the spatio-temporal relationships in multi-temporal remote sensing cloud removal. The core of this approach lies in constructing a model framework capable of deeply fitting the complex spatio-temporal dependencies between images from different time nodes. This explicit modeling of spatio-temporal relationships enables our method to more effectively capture the continuity and evolutionary patterns of surface features in the temporal dimension. Consequently, it achieves a more accurate restoration of cloud-free surface information in obscured regions while removing thick clouds, leading to a significant enhancement in the quality of the generated cloud-free imagery.

Our contributions are summarized as follows:

- We propose STAIT to explicitly model the inherent spatio-temporal correlations within multi-temporal data. STAIT innovatively employs an alternating iteration of spatial and temporal attention, enabling precise information aggregation from a global spatio-temporal context. Additionally, our approach offers a more direct and effective pathway to understanding and utilizing the dynamic characteristics of remote sensing imagery.
- To resolve the difficulties in feature tokens extraction and the excessive model complexity arising from high-dimensional multi-temporal inputs, we devise an efficient feature token generator. The generator leverages a group convolution-based multi-level structure. On the one hand, it utilizes group convolutions to reduce channel-wise computational redundancy and effectively control the model's parameter count. On the other hand, its multi-level design captures feature tokens at various scales, significantly strengthening the model's representation ability.
- To ensure high temporal consistency in the final cloud-free result, we introduce a novel pixel reconstruction decoder where the parameters of the upsampling module are shared across all timesteps. This constraint is critical as it forces the model to learn a unified generation style, effectively eliminating stylistic and textural inconsistencies.

The rest of this article is structured as follows. Section 2 reviews the related work. Section 3 elaborates on the proposed method. Section 4 reports and analyzes the experimental results, Section 5 discusses the findings and Section 6 provides the conclusion.

2. Related Work

2.1. Multi-Temporal Remote Sensing Image Cloud Removal

Existing approaches for multi-temporal remote sensing image cloud removal can be generally categorized into three groups: tensor completion via optimization algorithms, generative-model-based methods, and feature-learning-based methods. In the first category, the reconstruction of cloudy pixels is formulated as a tensor completion problem [14,17–23]. Within this framework, time-series images are typically represented as higher-order tensors, and optimization algorithms are utilized to recover cloudy values by exploiting the global low-rank structure inherent in the data. While mathematically rigorous, these methods often struggle with complex cloud patterns where the low-rank assumption may not fully hold. Furthermore, generative models represent another significant stream of research, principally including Generative Adversarial Networks (GANs) [24–26] and Denoising Diffusion Probabilistic Models (DDPMs) [27–29]. Researchers leverage adversarial training to synthesize cloud-free imagery or employ diffusion processes for gradual denoising, achieving excellent data distribution fitting. However, practical applications are often hindered by the training instability of GANs and the computational inefficiency associated with the iterative sampling of diffusion models.

More recently, feature-learning-based methods have gained prominence by leveraging deep neural networks to learn discriminative features and direct mappings from multi-temporal inputs to cloud-free outputs. This category predominantly employs Convolutional Neural Network (CNN) modules [30–33] to extract fine-grained local features or Transformer-based modules [34] to capture long-range contextual dependencies. By integrating their complementary strengths, these methods effectively suppress clouds and recover underlying surface information through an end-to-end learning paradigm.

2.2. Transformer in Remote Sensing Image

The evolution of deep learning in remote sensing largely mirrors advancements in general computer vision. Initially, CNNs dominated the field; however, their performance is often bottlenecked by limited receptive fields, which hinders the modeling of long-range dependencies. While architectures such as Feature Pyramid Networks (FPNs) [35], U-Net++ [36] and dense skip connections [37] were developed to fuse multi-scale features and mitigate information loss, they remain inherently limited in modeling long-range dependencies due to the local nature of convolution operations. Consequently, the field has witnessed a shift toward Transformers, which excel at global context modeling. Currently, the most common strategy involves hybrid architectures that combine CNNs for local feature extraction with Transformers for global representation. These hybrid approaches typically employ strategies such as series-parallel splicing [38] or multi-level hybridization [39]. Considering the rapid development of large-scale foundation models, Transformers are poised to become a significant trend in the next generation of RS image interpretation.

3. Methods

3.1. Problem Formulation and Motivation

The task of multi-temporal remote sensing cloud removal aims to reconstruct a cloud-free remote sensing image by leveraging multi-spectral cloudy images acquired from satellites at different time nodes. We define the multi-spectral cloudy images at various time nodes as $\mathbf{X} \in \mathbb{R}^{T\times C\times H\times W}$, where T represents the number of time nodes, C is the number of spectral channels, and H and W denote the height and width of the image, respectively. Specifically, $\mathbf{X}(T = i) \in \mathbb{R}^{C\times H\times W}$ denotes the multi-spectral cloudy image at a single time node i. Let $\mathbf{Y} \in \mathbb{R}^{C\times H\times W}$ and $\widehat{\mathbf{Y}} \in \mathbb{R}^{C\times H\times W}$ denote the ground-truth and the model-predicted cloud-free remote sensing images, correspondingly. Assuming that land cover features remain unchanged over a short period, our objective is to restore the underlying cloud-free remote sensing image $\mathbf{Y}$ from $\mathbf{X}$.

For a given region, we classify degraded pixels from the multi-temporal cloudy image sequence into two types, as illustrated in Figure 1. The first type is *Temporally Compensated Pixels*. These pixels are degraded in some observations, while they are clear (i.e., not degraded) in at least one temporal observation. The second type consists of *Blind-Inpainting Pixels*. These pixels are consistently degraded across all temporal observations (i.e., they are perpetually obscured by clouds or shadows). The *Temporally Compensated Pixels* are restored by incorporating prior information gleaned from their corresponding clear observations at different time nodes. In contrast, the *Blind-Inpainting Pixels* require the model to infer their values by learning the intrinsic prior distributions or structural regularities within remote sensing images.

Our approach strategically integrates both inter-temporal and spatial mechanisms: the former effectively handles *Temporally Compensated Pixels*, while the latter specifically targets *Blind-Inpainting Pixels*. Specifically, for *Temporally Compensated Pixels*, we devise an inter-temporal feature attention mechanism named Cross-Timenode Temporal Attention Module (CT-TAM) to capture the correlations between image features at different time nodes. For *Blind-Inpainting Pixels*, a spatial feature attention mechanism named Single-Timenode Spatial Attention Module (ST-SAM) is employed to learn the inherent prior feature distributions of remote sensing imagery, such as self-similarity. Finally, to reconstruct these two distinct categories of pixels, we propose a novel STAIT Module, which restores both pixel types in an alternating iterative manner. This iterative design is built upon a synergistic relationship. In each iteration, pixels restored using temporal information provide a reliable spatial context for the cloudy inpainting task. Consequently, the

newly inpainted regions create a more complete feature token, which in turn enhances the accuracy of temporal feature alignment in the subsequent iteration. This reciprocal process allows for the progressive refinement of the entire image. Notably, the decision to restore these two pixel types at the feature level is a key advantage of our approach. It allows the model to leverage higher-level semantic information, rather than manipulating raw pixel values directly. This enables the generation of more semantically coherent and visually plausible results, which is particularly crucial for reconstructing the *Blind-Inpainting Pixels*.

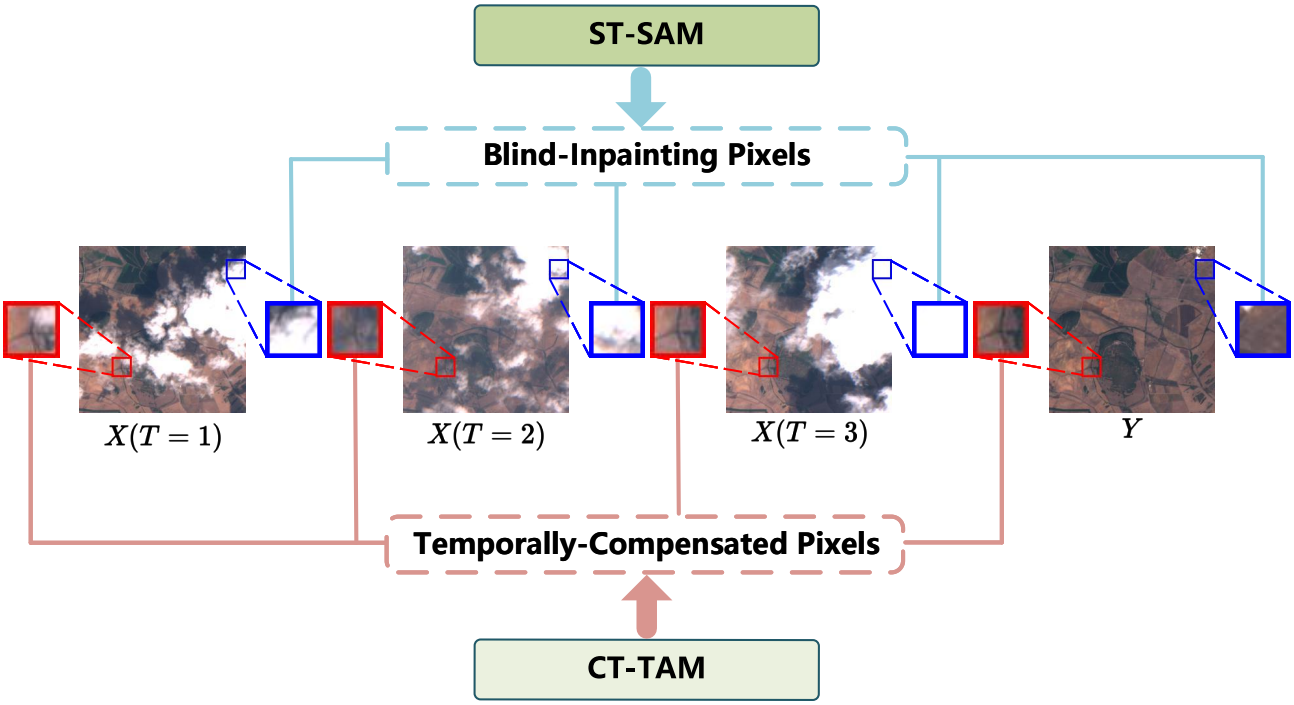


Figure 1. Categorization of degraded pixels in multi-temporal cloud removal into *Temporally Compensated* and *Blind-Inpainting types*.

3.2. Overall Framework

As illustrated in Figure 2, the proposed framework in this study is designed to synthesize high-quality cloud-free image from multi-temporal cloudy images (comprising T time nodes). The overall model architecture primarily comprises three core components: Feature Token Generator, STAIT Module, and Pixel Reconstruction Decoder.

Specifically, the processing pipeline unfolds as follows: Firstly, given a set of multi-temporal remote sensing cloudy images tensor $\mathbf{X}$. This tensor is then processed by the Feature Token Generator to produce a sequence of informative feature tokens. Subsequently, the output of the feature extraction module serves as the input to the STAIT module. The STAIT module's core lies in its unique alternating iterative mechanism, which, through iterative operations between the ST-SAM and CT-TAM, enables the deep excavation and fusion of intricate spatio-temporal patterns across different time nodes. Ultimately, the STAIT module outputs a rich spatio-temporal context-aware feature token representation. Finally, these features processed by the STAIT module are fed into the Pixel Reconstruction Decoder. This decoder is responsible for progressively reconstructing feature tokens into a high-resolution pixel image space, thereby synthesizing the final predicted cloud-free image.

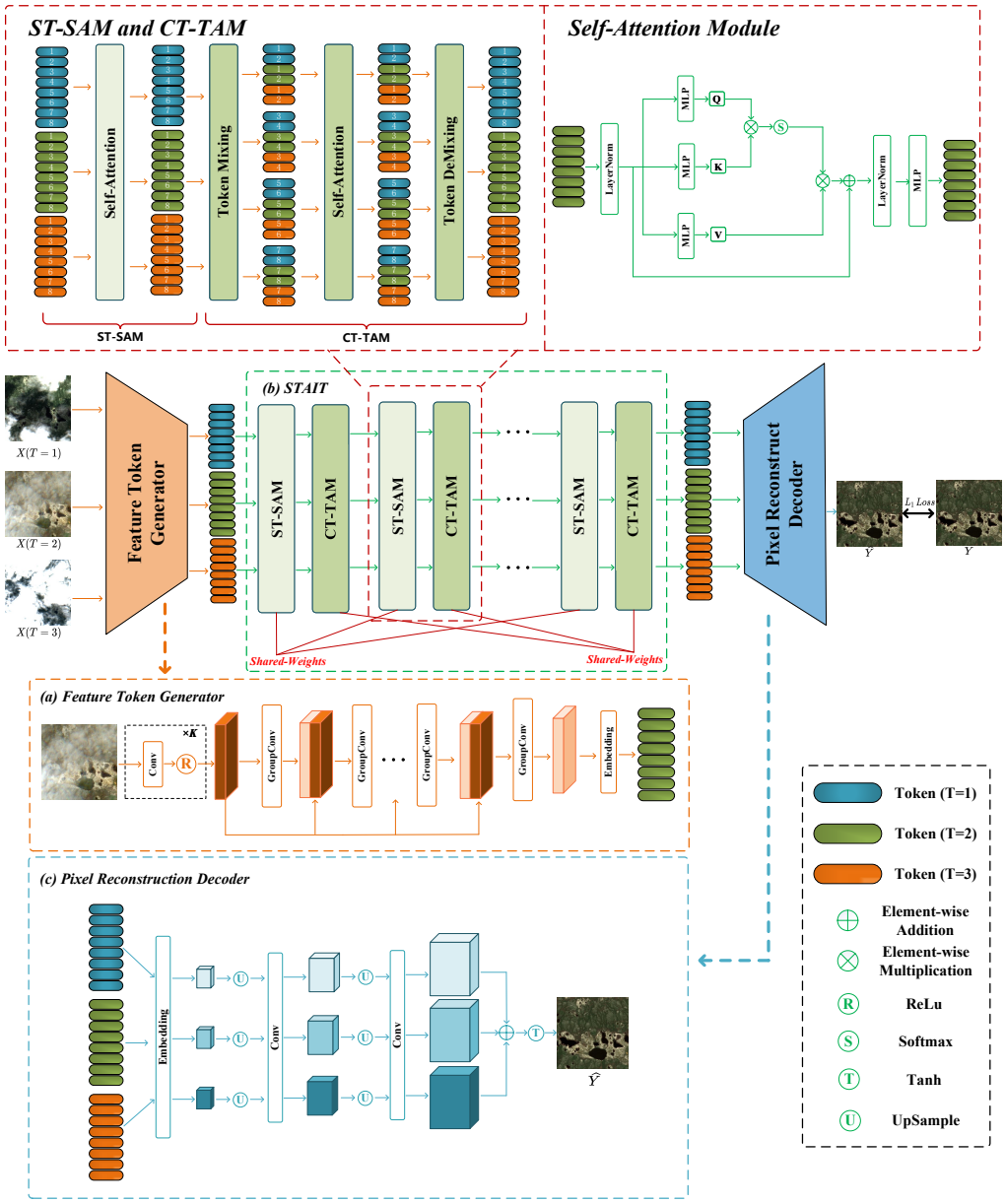


Figure 2. The overall architecture of our proposed network for multi-temporal cloud removal. The framework comprises three main stages: (**a**) a Feature Token Generator that encodes the multi-temporal cloudy images into feature tokens; (**b**) the core STAIT module that iteratively fuses spatial and temporal information, in this figure, we set $P = 4$ as an example; and (**c**) a Pixel Reconstruction Decoder that synthesizes the final cloud-free image from the refined tokens.

3.3. Feature Token Generator

A significant challenge in applying deep networks to multi-temporal data is the prohibitive computational cost associated with processing long image sequences. To mitigate this limitation, we propose an efficient feature token extraction strategy that employs two key components. Firstly, we use a multi-level feature extraction module to capture and fuse diverse features from shallow, intermediate, and deep layers, ensuring a comprehensive feature representation. Secondly, we utilize group convolution to reduce computational complexity. This dual approach enables both high computational efficiency and superior cloud removal performance. Specifically, for a given cloudy image $\mathbf{X}$, we first employ multi-layer convolutions to extract an initial feature, denoted as $\mathbf{F_0}$. Additionally, for input feature $\mathbf{F}$, the grouped convolution operation is denoted as G parallel standard convolutions:

$$\text{GroupConv}(\mathbf{F}) = \underset{g=1,\ldots,G}{\text{Concat}}\left(\text{Conv}(\mathbf{F}^g, \mathbf{W}^g)\right),$$

where $\mathbf{F}^g \in \mathbb{R}^{T\times(C/G)\times H\times W}$ is the g-th partition of $\mathbf{X}$ along the channel dimension, $\mathbf{W}^g$ is the g-th partition of $\mathbf{W}$ (corresponding to the filters for the g-th group), and Concat denotes concatenation along the output channel dimension. Ultimately, our proposed group convolution-based multi-level feature token extraction is represented as follows:

$$\begin{aligned}\widehat{\mathbf{F}}_{i+1} &= \text{GroupConv}(\mathbf{F}_i),\\ \mathbf{F}_{i+1} &= \text{Concat}(\widehat{\mathbf{F}}_{i+1}, \mathbf{F}_0),\\ \mathbf{F}_{gen} &= \mathbf{F}_N \in \mathbb{R}^{T\times C_{gen}\times H_{gen}\times W_{gen}},\end{aligned}$$

where $i = 0, \ldots, N-1$, C_{gen}, H_{gen} and W_{gen} represent the channel, height and width of feature $\mathbf{F}_{gen}$. At each level i, the features $\mathbf{F}_i$ undergo a group convolution operation to produce $\widehat{\mathbf{F}}_{i+1}$, capturing more abstract or discriminative representations. Crucially, these refined features $\widehat{\mathbf{F}}_{i+1}$ are then concatenated with the inital feature $\mathbf{F}_0$ to form the input for the next level $\mathbf{F}_{i+1}$. This repetitive concatenation of deeper, processed features with the foundational input $\mathbf{F}_0$ ensures that the network continuously benefits from both deep features semantic information and fine-grained shallow features details.

The cascaded multi-level feature extraction module significantly enhances feature extraction capabilities. By progressively integrating features from different abstraction levels while preserving the original input, it effectively mitigates information loss that may occur in deep networks. This hierarchical and recurrent feedback mechanism allows the module to capture a richer and more comprehensive representation of the input, leading to more robust and discriminative feature tokens.

3.4. Spatio-Temporal Alternating Iterative Transformer

To facilitate the robust reconstruction of occluded pixel regions, our methodology operates at the feature level, explicitly employing attention mechanisms to model inter-feature correlations. This application allows for the granular capture of two critical types of relationships: temporal dependencies among corresponding pixel feature tokens across varying time nodes, and spatial relationships between disparate pixel feature tokens within a given temporal period. By explicitly modeling these distinct contextual tokens, our system can more effectively infer and reconstruct missing or corrupted pixel information, leading to enhanced recovery fidelity.

We propose a novel Spatio-Temporal Alternating Iterative Transformer (STAIT) module designed to efficiently capture complex spatio-temporal dependencies within multi-temporal cloudy images. The core of this module lies in its unique alternating architecture, iteratively composed of a Single-Timenode Spatial Attention Module (ST-SAM) module and a Cross-Timenode Temporal Attention Module (CT-TAM) module. This design enables the simultaneous learning of rich and effective representations from both temporal and spatial dimensions. Specifically, given the input features $\mathbf{F}_{gen}$ obtained from a feature extraction layer, we first project them into a higher-dimensional feature space via an embedding layer, yielding $\mathbf{F}_{emb} \in \mathbb{R}^{T\times c\times d\times d}$. This embedding operation aims to enhance the expressive power of the features and adapt them to the internal dimensionality requirements of

the Transformer module. We then transform $\mathbf{F}_{emb}$ by unfolding it into the input tokens $\mathbf{M}(T=1), \mathbf{M}(T=2)$ and $\mathbf{M}(T=3) \in \mathbb{R}^{d^2 \times c}$ for the self-attention mechanism:

$$\mathbf{M}(T=i) = \text{UnFold}[(\mathbf{F}_{emb}(T=i)], \quad i=1,2,3,$$
$$\mathbf{M} = \underset{i=1,2,3}{\text{Concat}}[\mathbf{M}(T=i)], \quad \mathbf{M} \in \mathbb{R}^{T \times d^2 \times c},$$

where each row of $\mathbf{M}_i$ corresponds to a token's representation.

Subsequently, $\mathbf{M}$ is fed into our spatio-temporal alternating iterative process. We initialize $\mathbf{M}^0_{trans} = \mathbf{M}$ and for each iteration, with $l = 0, 1, \ldots, L-1$, the following steps are performed:

- **ST-SAM Step:** The primary function of this module is to learn correlations among different feature dimensions within a single timestep. By computing and aggregating this spatial contextual information, the module generates the feature representation $\mathbf{M}^{l+\frac{1}{2}}_{trans}$. This ensures that the network can comprehend and integrate the interactions between various tokens within each specific timestep. The process is described by the following equation:
$$\mathbf{M}^{l+\frac{1}{2}}_{trans} = \text{ST-SAM}_\theta(\mathbf{M}^l_{trans}),$$
- **CT-TAM Step:** Immediately following, $\mathbf{M}^{l+\frac{1}{2}}_{trans}$ is passed to the Cross-Timestep Temporal Attention Module. This module focuses on capturing correlations between tokens at the same spatial location but across different timesteps. By mixing different time nodes' tokens, CT-TAM can identify and learn patterns and dependencies of feature evolution over time, thereby yielding $\mathbf{M}^{l+1}_{trans}$. This process is also described by the following equation:
$$\mathbf{M}^{l+1}_{trans} = \text{CT-TAM}_\phi(\mathbf{M}^{l+\frac{1}{2}}_{trans}),$$

where θ and ϕ represent the learnable parameters of the ST-SAM and CT-TAM modules, respectively. *Notably, the parameters for both modules are shared across all steps of the alternating iteration process.*

This spatial-temporal attention alternation constitutes a complete iteration cycle. To thoroughly extract deep spatio-temporal patterns from the data, we repeat the aforementioned process L times. In each iteration, the output from the previous iteration serves as the input for the current one, allowing the network to progressively refine and strengthen its understanding of complex relationships across temporal and spatial dimensions, ultimately producing highly refined feature representations $\mathbf{M}^L_{trans}$. The core advantage of this alternating attention mechanism lies in its ability to systematically decouple and integrate spatio-temporal information flows. By processing these two dimensions independently and sequentially, our network mitigates potential dimensional confounding that might arise from traditional monolithic attention mechanisms, thereby enabling more precise and effective learning of both global and local spatio-temporal features. Subsequently, we define the self-attention mechanism as follows:

$$\text{Attention}(\mathbf{M}_i) = \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}})\mathbf{V},$$
$$\mathbf{Q} = \mathbf{M}_i\mathbf{W}_\mathbf{Q}, \mathbf{K} = \mathbf{M}_i\mathbf{W}_\mathbf{K}, \mathbf{V} = \mathbf{M}_i\mathbf{W}_\mathbf{V},$$

where $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V} \in \mathbb{R}^{c \times d^2}$ denote the learnable Query, Key, and Value matrices, respectively, obtained by applying linear transformations $\mathbf{W}_\mathbf{Q}$, $\mathbf{W}_\mathbf{K}$, and $\mathbf{W}_\mathbf{V} \in \mathbb{R}^{d \times d}$. The architectures of the ST-SAM and CT-TAM modules, illustrated in Figure 2, are introduced as follows:

Single-Timenode Spatial Attention Module. The primary function of the ST-SAM module is to enhance feature representations for the inpainting of degraded regions. It operates by exclusively leveraging the spatial dependencies among feature tokens within each individual time step. Given an input feature tensor $\mathbf{M}_{trans}^{l}$ to the SAM, a shared self-attention mechanism is employed to compute the internal feature similarities for each time node. This process enables the module to capture the contextual dependencies among spatial elements within a single time node, thereby optimizing the per-time-nodes feature representation prior to subsequent temporal processing. The process can be formulated as:

$$\begin{aligned}
&\mathbf{M}_{trans}^{l}(T=1), \mathbf{M}_{trans}^{l}(T=2), \mathbf{M}_{trans}^{l}(T=3) = \text{Split}[\mathbf{M}_{trans}^{l}], \\
&\mathbf{M}_{trans}^{l+\frac{1}{2}} = \underset{i=1,2,3}{\text{Concat}}[\text{Attention}_{\theta}(\mathbf{M}_{trans}^{l}(T=i))].
\end{aligned}$$

Cross-Timenode Temporal Attention Module. The CT-TAM module is designed to capture correlations among feature tokens in the same spatial regions across different time nodes. Given an input feature token $\mathbf{M}_{trans}^{l+\frac{1}{2}}$, we first partition all tokens into P distinct segments for each time node. Subsequently, for each segment index $p \in \{1, 2, \ldots, P\}$, we concatenate the p-th segment from all time nodes to form $\mathbf{m}_{trans}^{l+\frac{1}{2}}(p=i) \in \mathbb{R}^{(\frac{d^2}{P \times T}) \times c}$. This step embodies the core mechanism of our approach, which we denote as token mixing. A self-attention mechanism is then applied to these mixed feature tokens $\mathbf{m}_{trans}^{l+\frac{1}{2}}$, enabling the module to compute the cross-temporal relationships for that specific area. Finally, the inverse of the token mixing operation named token demixing is performed to restore the token representation to the original input state. This process effectively captures the relationships between features in the same spatial regions across varying time nodes. The process can be formulated as:

$$\begin{aligned}
&\mathbf{m}_{trans}^{l+\frac{1}{2}}(p=1), \mathbf{m}_{trans}^{l+\frac{1}{2}}(p=2), \cdots, \mathbf{m}_{trans}^{l+\frac{1}{2}}(p=P) = \text{Token Mixing}[\mathbf{M}_{trans}^{l+\frac{1}{2}}], \\
&\mathbf{M}_{trans}^{l+1} = \text{Token Demixing}\{\underset{i=1,2,\ldots,P}{\text{Concat}}[\text{Attention}_{\phi}(\mathbf{m}_{trans}^{l+\frac{1}{2}}(p=i)]\}.
\end{aligned}$$

3.5. Pixel Reconstruction Decoder

To generate the final cloud-free image, we employ a Pixel Reconstruction Decoder to transform the feature tokens from the STAIT module. Initially, these tokens are adapted via a Multi-Layer Perceptron (MLP) and a reshape operation to match the decoder's input dimensions. The tokens are then processed by a shared upsampling module, which progressively increases their resolution through alternating upsampling and convolutional layers until they reach the target image size. Critically, the parameters of this module are shared across all timesteps, a constraint that forces the model to learn a unified generation style and ensures strong stylistic and textural consistency in the outputs. The final prediction is produced by summing the reconstructed images from each timestep and applying a non-linear activation function to effectively aggregate the multi-temporal information.

3.6. Loss Function

The proposed network is optimized using an ℓ_1 loss function. This function computes the mean absolute error (MAE) between the cloud-free image $\mathbf{Y} \in \mathbb{R}^{C \times H \times W}$ and the prediction $\widehat{\mathbf{Y}} \in \mathbb{R}^{C \times H \times W}$, as shown in the following formula:

$$\mathcal{L}(\widehat{\mathbf{Y}}, \mathbf{Y}) = \frac{1}{C \times H \times W} \sum_{c=1}^{C} \sum_{i=1}^{H} \sum_{j=1}^{W} \left| \widehat{\mathbf{Y}}_{c,i,j} - \mathbf{Y}_{c,i,j} \right|, \tag{1}$$

where C, H, and W denote the number of spectral bands, height, and width, respectively. In our implementation, the loss is computed as the mean reduction over all spectral dimensions (bands) and spatial dimensions (pixels). The ℓ_1 loss function is particularly effective due to its robustness to outliers, which prevents large errors from dominating the optimization.

4. Experiments

To validate the effectiveness of our proposed model, we conduct a series of comprehensive experiments in this section. We first compare our method with several state-of-the-art approaches on two benchmark datasets to validate its performance. Furthermore, we conduct extensive ablation studies to analyze the contribution of each key component, thereby justifying the rationale behind our design.

4.1. Datasets and Metrics

To comprehensively evaluate the performance of our proposed method, we conducted experiments on two public multi-temporal remote sensing cloud removal benchmark datasets: STGAN [24] and Sen2_MTC_New [40]. All images in these datasets have a spatial resolution of 256×256 pixels and 3 time nodes, which means $T = 3$. The details of these datasets are as follows:

- STGAN is a multispectral dataset derived from the Sentinel-2 satellite. It contains 3130 image sequences from different geographical regions. Each sequence is composed of three cloudy images and a corresponding cloud-free reference image. All images include four spectral bands: Red, Green, Blue (RGB), and Near-Infrared (NIR). In our experiments, we adopted a random splitting strategy, partitioning the dataset into training, validation, and test sets at an 8:1:1 ratio.
- Sen2_MTC_New is structurally similar to STGAN, comprising a total of 3417 multispectral image sequences. Each sequence also consists of three cloudy images, one cloud-free reference image, and four spectral bands. To ensure a fair comparison with previous works, we strictly adhered to the standard partitioning scheme proposed in [40] to construct the training, validation, and test sets.

To quantitatively evaluate the quality of the generated cloud-free images, we adopted three widely used evaluation metrics: Peak Signal-to-Noise Ratio (PSNR), the Structural Similarity Index [41] (SSIM) and Spectral Angle Mapper (SAM).

PSNR is a common metric for measuring the quality of image reconstruction. It assesses fidelity by calculating the pixel-level error between the generated image and the ground-truth reference image. A higher PSNR value indicates less distortion and, therefore, higher quality in the reconstructed image. Its formula is defined as:

$$\text{PSNR} = 10 \cdot log_{10}(\frac{\text{MAX}_I^2}{\text{MSE}}),$$

where MAX_I^2 is the maximum possible pixel value of the image, and MSE is the Mean Squared Error between the two images $\widehat{\mathbf{Y}}$ and $\mathbf{Y}$.

SSIM evaluates image quality from the perspective of the human visual system by comparing the similarity of two images in terms of three components: luminance, contrast, and structure. The SSIM value ranges from -1 to 1, where a value closer to 1 signifies greater structural similarity between the two images and a better generation result. It is defined as:

$$\text{SSIM}(\widehat{\mathbf{Y}}, \mathbf{Y}) = \frac{(2\mu_{\widehat{\mathbf{Y}}}\mu_{\mathbf{Y}} + c_1)(2\sigma_{\widehat{\mathbf{Y}}\mathbf{Y}} + c_2)}{(\mu_{\widehat{\mathbf{Y}}}^2 + \mu_{\mathbf{Y}}^2 + c_1)(\sigma_{\widehat{\mathbf{Y}}}^2 + \sigma_{\mathbf{Y}}^2 + c_2)},$$

where $\widehat{\mathbf{Y}}$ and $\mathbf{Y}$ represent the generated and reference images, respectively; μ and σ are the mean and standard deviation; and c_1, c_2 are stabilization constants to avoid division by zero.

The Spectral Angle Mapper (SAM) is a metric used to determine the spectral similarity between two spectra by calculating the angle between them. It treats each spectrum as a vector in an L-dimensional space, where L is the number of spectral bands. This makes SAM insensitive to illumination changes, as it measures the similarity of the spectral shape rather than magnitude. A smaller SAM value indicates a higher similarity. The formula is:

$$\mathrm{SAM}(\widehat{\mathbf{Y}}, \mathbf{Y}) = \arccos\left(\frac{\widehat{\mathbf{Y}}^T\mathbf{Y}}{\|\widehat{\mathbf{Y}}\|_2\|\mathbf{Y}\|_2}\right),$$

where $\mathbf{Y}$ is the ground-truth spectral vector and $\widehat{\mathbf{Y}}$ is the estimated spectral vector. The term $\|\cdot\|_2$ denotes the ℓ_2-norm.

4.2. Implementation Details

The proposed model was implemented in PyTorch 2.2.1 and trained on an NVIDIA RTX 4090 GPU. Key architectural hyperparameters K, N, P and L were all set to 4. The embedding dimension d is set to 256. For network optimization, the Adam optimizer was employed with a learning rate of 1×10^{-4}, and coefficients β_1 and β_2 were set to 0.9 and 0.99, respectively. The model was trained for 200,000 iterations with a batch size of 16. For the evaluation, the model checkpoint from the final training epoch was utilized.

4.3. Experiments on STGAN Dataset

To conduct a rigorous evaluation of our proposed method, we adopted the publicly available dataset and the official training, validation, and test splits from STGAN [24]. This standardized protocol ensures both the reproducibility of our experiments and the fairness of the comparison. Our method was comprehensively benchmarked against the traditional temporal Median Filter, which generates the output by taking the median of the multi-temporal inputs, and a suite of state-of-the-art (SOTA) multi-temporal deep learning cloud removal techniques, including DSen2-CR [42], CTGAN [40], STGAN [24], CR-TS-Net [43], PMAA [44] and UnCRtainTS [31]. Regarding the baseline comparisons, we adopt a hybrid strategy to ensure each method achieves its optimal performance. For methods where only training code is provided, we retrain the models from scratch using our dataset partitions. For methods with officially released pre-trained weights that utilize the same dataset partitions as ours, we directly employ the official models to avoid performance degradation caused by sub-optimal retraining. To maintain consistency, all methods are evaluated on the identical testing set using unified metrics. To ensure an equitable comparison, all baseline models were retrained from scratch on the designated training set until convergence, utilizing their official source code and adhering strictly to the hyperparameter configurations recommended in their respective original publications.

The qualitative comparison results are presented in Figure 3, which illustrates the cloud removal performance of different methods across several representative scenes. It is evident that most baseline methods exhibit varying degrees of artifacts and limitations, particularly when processing large areas of thick cloud or the fine details at the edges of thin clouds. For instance, the results generated by compared methods suffer from conspicuous visual artifacts and textural distortions. In stark contrast, our proposed method demonstrates significant superiority in visual quality. Our method effectively preserves the spectral characteristics of the original land cover. The resulting images exhibit natural and coherent colors that are seamlessly integrated with the surrounding context. These

advantages underscore our model's robust capability to effectively capture and leverage multi-temporal information for the reconstruction of high-quality, cloud-free pixels.

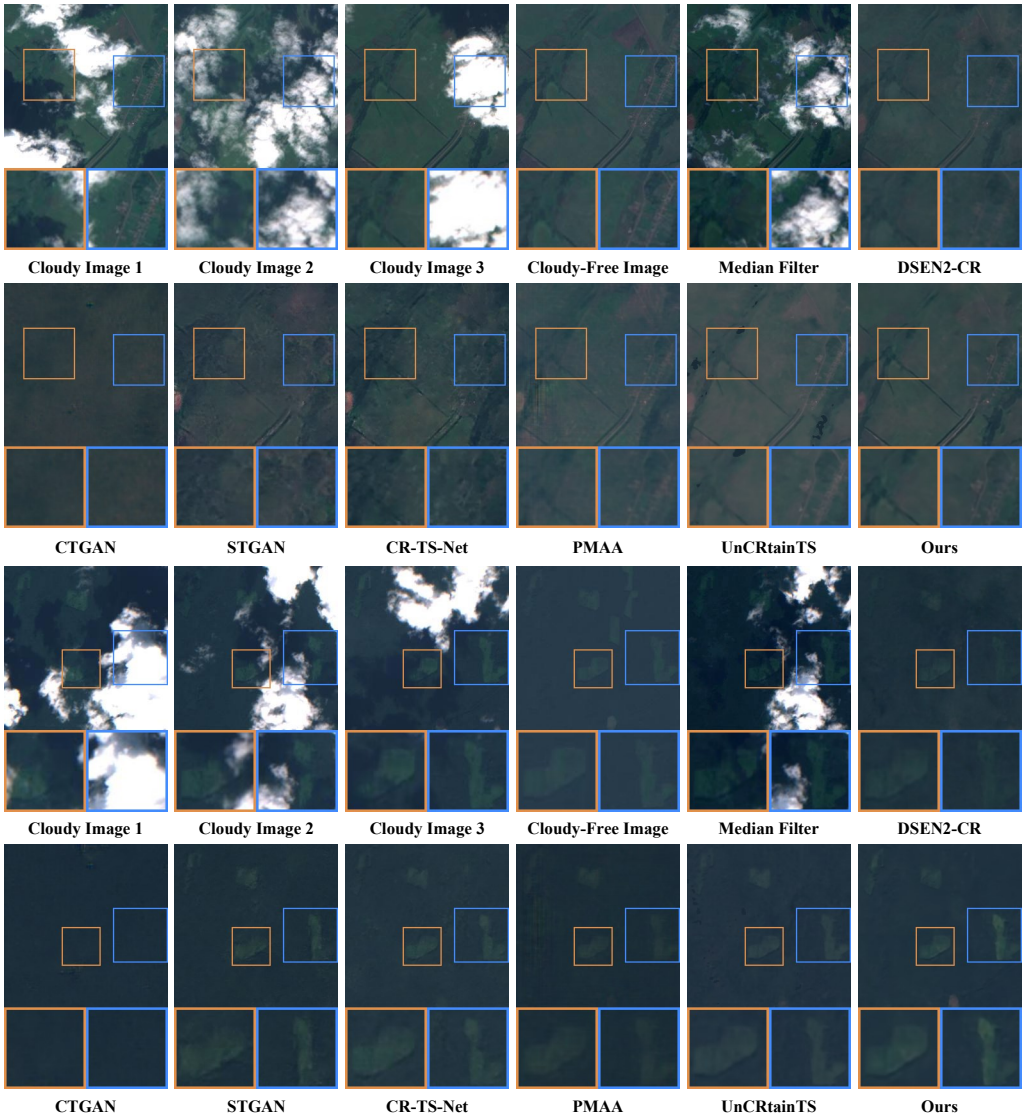


Figure 3. Qualitative comparison on the STGAN dataset. The top two rows show results for the "40UCG_20008000" scene, and the bottom two rows show results for the "37VCG_60009000" scene. The highlighted regions (blue and orange) are magnified for clarity.

As delineated in Table 1, our method achieves the highest scores across both key metrics. Specifically, our model obtains a PSNR of 27.340 and an SSIM of 0.831, outperforming all other competing methods by a significant margin. The top-ranking SSIM score corroborates our qualitative findings, confirming that our method not only ensures pixel accuracy but also excels at preserving the intrinsic structure, luminance, and contrast of the original ground features. This aspect is of paramount importance for downstream remote sensing applications, such as change detection and land cover classification, which are highly dependent on the structural integrity of the imagery. To evaluate the stability and reproducibility of our method, we conducted several independent runs with different random seeds. The proposed method demonstrates robust performance with a mean PSNR of 27.321 $\pm$ 0.121 dB, SSIM of 0.827 $\pm$ 0.014, and SAM of 2.946 $\pm$ 0.018. The low standard deviations indicate that our model converges consistently to a high-quality image.

In summary, the convergence of evidence from both qualitative visual comparisons and quantitative metric evaluations consistently demonstrates that our proposed multi-temporal cloud removal method surpasses current mainstream approaches in overall performance. Our model not only generates cloud-free images that are visually more realistic, richer in detail, and more color-consistent, but its superior reconstruction accuracy and structural fidelity are also empirically validated by objective metrics. These comprehensive results strongly affirm the efficacy and advanced nature of our designed model.

Table 1. Quantitative comparison results on STGAN and SEN2_MTC_New datasets. The red and blue markers represent the best and second-best values.

Methods	STGAN Dataset			SEN2_MTC_New Dataset		
	PSNR ↑	SSIM ↑	SAM ↓	PSNR ↑	SSIM ↑	SAM ↓
Median Filter	10.495	0.441	7.756	9.485	0.404	11.658
DSen2-CR	25.559	0.788	3.801	17.417	0.576	7.378
CTGAN	25.211	0.776	4.079	18.308	0.609	6.857
STGAN	26.310	0.796	3.376	18.158	0.556	7.018
CR-TS-Net	26.309	0.800	3.346	18.597	0.616	6.892
PMAA	26.930	0.829	3.269	18.369	0.614	7.155
UnCRtainTS	27.244	0.821	3.112	18.770	0.631	6.528
STAIT (Ours)	27.340	0.831	2.949	18.868	0.640	6.392

4.4. Experiments on SEN2_MTC_New Dataset

We conducted additional experiments on the SEN2_MTC_New dataset. We adhered to the identical experimental protocol established for the STGAN dataset to ensure a consistent and fair comparison. For the baseline methods, it is noteworthy that we directly employed the publicly available, pre-trained models for PMAA and CTGAN. All other competing methods were retrained from scratch on the SEN2_MTC_New training set until convergence.

Figure 4 provides a visual testament to our model's performance on the more challenging SEN2_MTC_New dataset. The comparative results reveal the limitations of existing methods when faced with diverse and complex cloud scenarios. Many competing models, such as DSen2-CR and STGAN, produce reconstructions that either degrade the visual plausibility with unnatural textures, leading to a "patchwork" effect where the restored area is clearly demarcated. Conversely, our approach yields reconstructions that are markedly more plausible and coherent. Our model excels in rendering fine-scale details with high precision. The generated content maintains spatial coherence and is semantically consistent with the surrounding cloud-free context. A key strength of our method is its ability to achieve seamless tonal and color blending. The restored regions are photorealistically integrated into the image, avoiding the chromatic aberrations that plague other methods. This visual evidence suggests that our model possesses a superior understanding of multi-temporal scene dynamics, enabling it to generate restorations that are not only free of clouds but also visually indistinguishable from genuine observations.

To substantiate these visual observations with empirical data, we performed a quantitative comparison using the PSNR and SSIM metrics. The numerical results, summarized in Table 1, unequivocally demonstrate the quantitative superiority of our proposed method. Our model consistently achieves the highest scores, establishing a clear performance advantage over all evaluated SOTA techniques on the SEN2_MTC_New dataset. The top-tier PSNR score signifies a greater minimization of pixel-level discrepancies between our restored images and the ground truth. This highlights our model's capacity for precise and accurate pixel synthesis. The leading SSIM score is particularly telling, as it reflects our model's proficiency in preserving the complex structural information and perceptual

quality of the original scene. This is crucial for ensuring that the restored images are not just mathematically close but also visually faithful. The consistent lead in both metrics underscores the robustness and effectiveness of our architectural design, validating its ability to generalize to new and varied data distributions. Similarly, we applied the same rigorous evaluation protocol to the SEN_MTC_New dataset. By performing independent runs with distinct random seeds, our method achieved a mean PSNR of 18.871 $\pm$ 0.124 dB, SSIM of 0.642 $\pm$ 0.015, and SAM of 6.384 $\pm$ 0.182. These consistent results with minimal standard deviations further confirm the stability of our architecture.

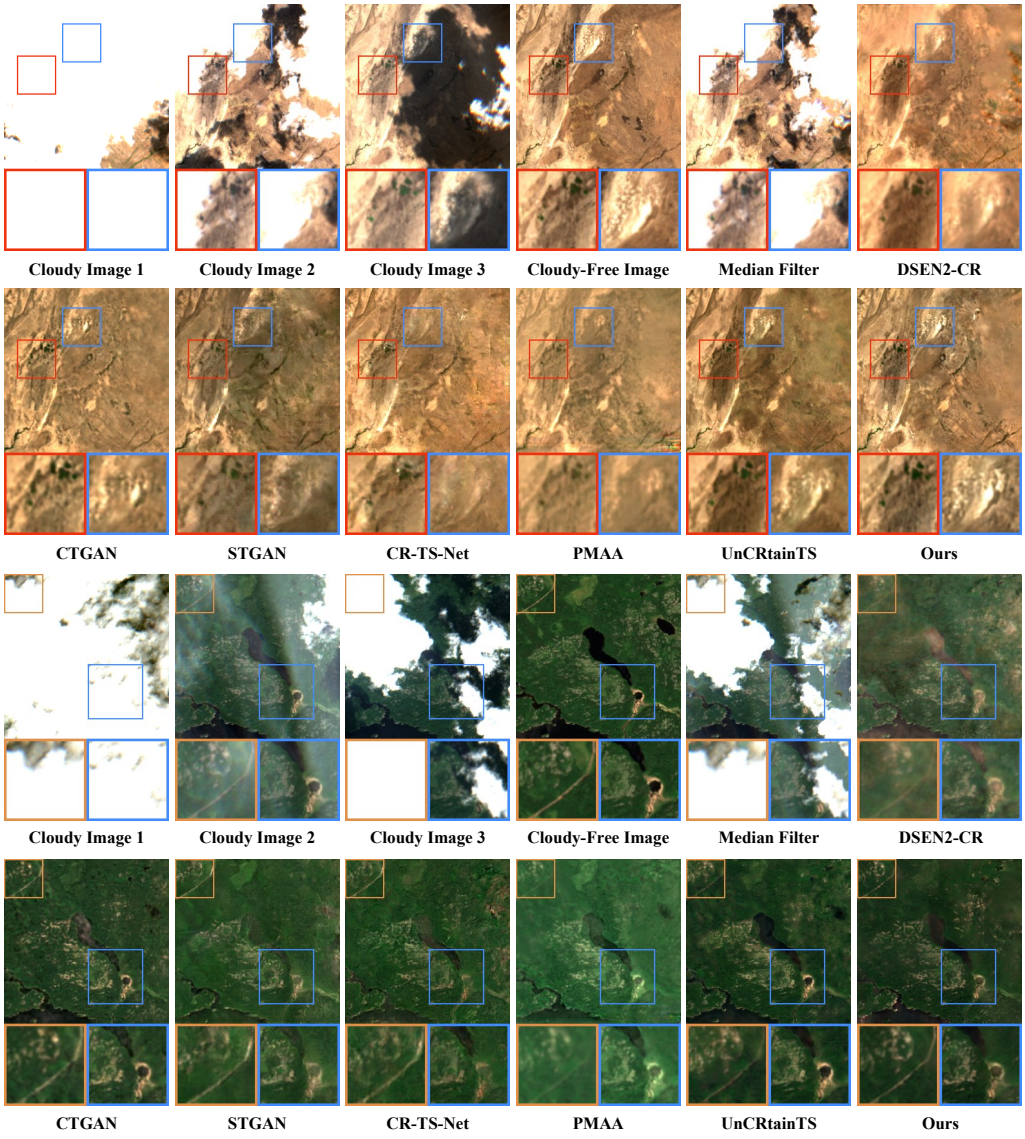


Figure 4. Qualitative comparison on the SEN2_MTC_New dataset. The top two rows show results for the "T41UNR_R020_57" scene, and the bottom two rows show results for the "T15UVU_R012_43" scene. The highlighted regions (blue, red and orange) are magnified for clarity.

To ensure a fair comparison with the state-of-the-art diffusion-based method DiffCR [27], we evaluated it using its officially released pre-trained weights on the standard test set. Since DiffCR is designed to operate on 3-channel RGB data, we correspondingly re-trained and tested our proposed STAIT framework under the same 3-channel configuration, which is denoted as STAIT-3. As presented in Table 2, although DiffCR achieves marginally higher PSNR and SSIM scores, our method outperforms it in terms of SAM, indicating superior spectral fidelity. More importantly, our method demonstrates a significant advan-

tage in computational efficiency. It is worth noting that all efficiency metrics (FLOPs and latency) were evaluated by a fixed input size of 256 × 256 with 3 time nodes. Compared to DiffCR, STAIT requires significantly fewer parameters and achieves an inference speed that is orders of magnitude faster (6.55 ms vs. 498.70 ms), making it far more practical for real-time applications.

Table 2. Comprehensive comparison of model complexity (Efficiency) and restoration quality (Performance) on the SEN2_MTC_New dataset.

Methods	Efficiency Metrics			Performance Metrics		
	Params (M)	FLOPs (G)	Latency (ms) ↓	PSNR ↑	SSIM ↑	SAM ↓
DiffCR	22.91	91.72	498.70	19.150	0.671	6.454
STAIT-3 (Ours)	7.68	118.75	6.55	18.880	0.653	6.324

Taken in concert, the qualitative and quantitative results on the SEN2_MTC_New dataset reinforce the conclusions drawn from our initial experiments. Our method sets a new performance benchmark for multi-temporal cloud removal. The generated images are characterized by a remarkable degree of realism, detail preservation, and color accuracy, which is corroborated by their state-of-the-art performance on objective metrics. This sustained success across different datasets highlights the robustness of our approach. It effectively addresses the core challenges of the task, demonstrating its potential as a reliable and highly effective tool for generating analysis-ready, cloud-free satellite imagery for critical downstream applications.

4.5. Efficiency Analysis

In addition to restoration quality, computational efficiency is a critical determinant for the practical deployment of cloud removal algorithms, particularly in real-time remote sensing scenarios. As presented in Table 3, our proposed STAIT demonstrates a significant advantage in computational costs. Evaluated on 4-channel inputs with a spatial resolution of 256 × 256 across 3 time nodes, STAIT achieves the lowest FLOPs of 118.81 G and an ultra-low inference latency of 6.68 ms. Regarding memory consumption, our method maintains a competitive footprint of 574.44 MB. While slightly higher than the parameter-sparse DSen2-CR, it remains significantly more memory-efficient than heavy-weight counterparts like CTGAN and CR-TS-Net. More importantly, STAIT represents a 2.5× speedup compared to the second-best method, PMAA, striking a superior balance between memory usage and processing speed. This efficiency is primarily attributed to our specific architectural designs: on the one hand, we utilize grouped convolutions to replace standard dense convolutions in high-dimensional feature extraction stages, which significantly reduces computational redundancy and FLOPs without compromising feature representation; on the other hand, the adoption of a shared-parameter decoder effectively constrains the growth of model parameters and minimizes memory access overhead during inference. Consequently, STAIT achieves an optimal trade-off among performance, latency, and memory costs, rendering it highly suitable for large-scale and real-time on-board processing tasks.

4.6. Analysis of Hyperparameters

In this section, we conduct a series of experiments to determine the optimal settings for our model's key hyperparameters. We individually assess the impact of: (1) the number of levels N in the Feature Extraction module, and (2) the number of iterations L in the STAIT module to find a balance between performance and computational cost.

Table 3. Params, FLOPs, Inference Time and Memory Comparison. The red and blue markers represent the best and second-best values.

Methods	Params, FLOPs, Latency and Memory Comparison			
	Params (M)	**FLOPs (G)**	**Latency (ms)**	**Memory (MB)**
DSen2-CR	18.92	2478.73	44.77	333.39
CTGAN	642.92	1263.40	43.20	3319.50
STGAN	231.93	2186.51	34.25	1353.24
CR-TS-Net	38.39	15082.58	312.30	1575.43
PMAA	3.45	185.09	16.42	370.59
UnCRtainTS	0.56	167.10	38.90	785.32
STAIT (Ours)	7.68	118.81	6.68	574.44

Firstly, we investigate the importance of the hierarchical feature extraction mechanism by varying the number of levels N, within the Feature Extraction module, setting N to 2, 4, 6, and 8. As presented in Table 4, the results exhibit a clear and positive correlation between the number of levels N and the model's performance when N = 2, 4. The steady improvement in both PSNR SSIM and SAM scores with increasing N demonstrates that a deeper feature hierarchy enables the model to capture a richer and more diverse set of spatial features, which is crucial for high-fidelity image reconstruction. It is observed that the performance peaks at N = 4. Secondly, we aim to quantify the effect of the iterative refinement process within our STAIT module. We configured the model with a varying number of alternating iterations L, specifically testing L = 2, 4, 6, and 8. The quantitative results detailed in Table 4 reveal that performance initially improves as L increases from 2 to 4, but subsequently degrades when L exceeds 4. This indicates that L = 4 strikes an optimal balance. Therefore, considering the trade-off between performance and computational complexity, we select N = 4 and L = 4 as the optimal setting for our model.

Table 4. Ablation analysis of hyperparameters N and L on the SEN2_MTC_New dataset. The red markers represent the best values.

Analysis of Hyperparameter N				Analysis of Hyperparameter L			
N	**PSNR ↑**	**SSIM ↑**	**SAM ↓**	***L***	**PSNR ↑**	**SSIM ↑**	**SAM ↓**
2	17.842	0.588	7.157	2	17.754	0.583	7.352
4	18.868	0.640	6.392	4	18.868	0.640	6.392
6	18.859	0.637	6.458	6	18.851	0.639	6.405
8	18.731	0.638	6.437	8	18.865	0.635	6.428

4.7. Ablation Study on STAIT Module

Having determined the optimal hyperparameters, we conducted a comprehensive ablation study to validate the architectural effectiveness of STAIT, ranging from the fundamental backbone choice to specific module designs. First and foremost, to substantiate the advantage of our iterative Transformer architecture over traditional convolutional methods, we established a baseline named "w/o STAIT," where the entire STAIT module (including ST-SAM and CT-TAM) was replaced by standard residual convolutional blocks of equivalent depth. As detailed in Table 5, this configuration suffered the most severe performance degradation, with a PSNR drop of 1.33 dB compared to the full model. This sharp decline underscores the limitation of standard CNNs in capturing the complex, long-range spatio-temporal dependencies required for cloud removal, thereby validating the necessity of the proposed attention-based architecture.

Table 5. Ablation Experiments on SEN2_MTC_New Dataset. The red markers represent the best values.

Ablation Experiment			
Exp	**PSNR ↑**	**SSIM ↑**	**SAM ↓**
w/o STAIT	17.535	0.541	7.662
CT-TAM only	18.112	0.577	7.395
ST-SAM only	17.631	0.538	7.928
w/o Feature Token Generator	18.331	0.614	6.883
w/o Shared Decoder	18.742	0.632	6.487
Ours	18.868	0.640	6.392

Subsequently, we dissected the internal structure of the STAIT module to demonstrate the synergistic contributions of its two core sub-modules: the ST-SAM and the CT-TAM. We constructed two homogeneous variants: "ST-SAM only," where the CT-TAM was replaced by a second ST-SAM to force exclusive reliance on spatial processing, and "CT-TAM only," where the ST-SAM was replaced by an additional CT-TAM to rely solely on temporal aggregation. As evidenced by the results, both configurations experienced significant performance drops compared to the heterogeneous full model. This outcome decisively demonstrates that ST-SAM and CT-TAM fulfill complementary and non-interchangeable roles. Neither module, even when duplicated, can effectively substitute for the other, confirming that the dynamic interplay between spatial texture refinement and temporal consistency aggregation is fundamental to STAIT.

Finally, we investigated the impact of specific component designs regarding feature embedding and decoding strategies. The experiment "w/o Feature Token Generator," which replaced our tokenization module with a standard dense CNN for feature extraction, resulted in a clear performance decline. This indicates that our token generator effectively encapsulates high-level semantic information into tokens, which are more suitable for subsequent Transformer processing than simple convolutional features. Furthermore, the "w/o Shared Decoder" variant, which utilized independent decoders for each time point instead of a shared-parameter strategy, also failed to match the optimal performance. This suggests that the shared-parameter design not only reduces model complexity but also acts as a regularization mechanism, implicitly enforcing consistency across temporal outputs. In summary, the superior performance of the full STAIT model confirms that each component is indispensable for optimal restoration results.

5. Discussion

In this study, our primary objective was to overcome the limitations of existing deep learning paradigms that fail to capture complex spatio-temporal dynamics in multi-temporal cloud removal. The experimental results demonstrate that the proposed Spatio-Temporal Alternating Iterative Transformer (STAIT) significantly outperforms traditional concatenation-based approaches. Unlike previous methods that often treat multi-temporal images as isolated inputs or simple channel-wise stacks, STAIT explicitly models the evolutionary relationships across temporal sequences. By employing an alternating mechanism between spatial and temporal attention, our model effectively disentangles latent true surface information from cloud interference. This explicit modeling ensures that the reconstructed regions are not only visually plausible but also maintain strong temporal consistency with the surrounding environment, validating our hypothesis that capturing dynamic dependencies is crucial for high-fidelity restoration.

Beyond reconstruction accuracy, the proposed framework successfully addresses the practical challenges of model complexity and output coherence. The integration of the group convolution-based image token generator plays a vital role in mitigating the com-

putational redundancy typically associated with high-dimensional multi-temporal inputs. This design allows for efficient multi-level feature extraction without compromising the model's representational capacity. Furthermore, the introduction of a shared progressive upsampling network in the decoder significantly enhances temporal coherence. By enforcing shared parameters across different timestamps, the model is constrained to learn a unified generation style. This effectively eliminates the textural and stylistic inconsistencies that frequently degrade the quality of results in prior feature-learning frameworks, ensuring that the cloud-free output remains natural and artifact-free.

While the proposed framework successfully optimizes computational efficiency by significantly reducing parameters and FLOPs, its performance remains constrained by the intrinsic information availability within the input sequence. The restoration of high-fidelity imagery relies heavily on extracting complementary cues from the temporal dimension. In scenarios characterized by extreme temporal intervals or persistent, extensive cloud cover across all available time points, the retrieval of fine-grained surface details becomes an ill-posed problem. Under such conditions, the lack of visible reference pixels and weak temporal correlations can impede the reconstruction of high-frequency textures. Consequently, future work will focus on addressing these data-dependent challenges, potentially by incorporating stronger external priors or generative constraints to enhance feature recovery in highly occluded or temporally sparse sequences.

6. Conclusions

To address the challenge of thick cloud removal in multi-temporal imagery, we propose a novel framework that effectively addresses the absence of explicit spatio-temporal dependency modeling, a critical limitation in existing methods. At the core of our proposed framework is the Spatio-Temporal Alternating Iterative Transformer (STAIT), a novel module that captures the dynamic evolution of surface features through an alternating application of spatial and temporal attention. Complemented by an efficient multi-level feature extractor and a weight-sharing reconstruction decoder, our framework effectively manages model complexity while ensuring high-fidelity and consistent image synthesis. Experimental results demonstrate that our method significantly improves the accuracy and visual quality of cloud-free images, offering a robust, parameter-efficient solution that marks a significant step forward in achieving continuous and reliable Earth observation.

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

Funding: This research was funded by the National Natural Science Foundation of China (Grants No. 12371512), Young Talent Fund of Xi'an Association for Science and Technology (Grants No. 0959202513207) and Research Fund of Yunnan Key Laboratory of Intelligent Systems and Computing under Grant ISC25Y02.

Data Availability Statement: The original data presented in the study are openly available from the following repositories: the STGAN dataset (https://dataverse.harvard.edu/citation?persistentId=doi:10.7910/DVN/BSETKZ, accessed on 10 February 2026) and the Sen2_MTC_New dataset (https://github.com/come880412/CTGAN, accessed on 10 February 2026).

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

References

1. Stumpf, F.; Schneider, M.K.; Keller, A.; Mayr, A.; Rentschler, T.; Meuli, R.G.; Schaepman, M.; Liebisch, F. Spatial monitoring of grassland management using multi-temporal satellite imagery. *Ecol. Indic.* **2020**, *113*, 106201. [CrossRef]
2. Thürkow, F.; Lorenz, C.G.; Pause, M.; Birger, J. Advanced Detection of Invasive Neophytes in Agricultural Landscapes: A Multisensory and Multiscale Remote Sensing Approach. *Remote Sens.* **2024**, *16*, 500. [CrossRef]
3. Cao, B.; Kang, L.; Yang, S.; Tan, D.; Wen, X. Monitoring the Dynamic Changes in Urban Lakes Based on Multi-source Remote Sensing Images. In *Communications in Computer and Information Science, Proceedings of the Geo-Informatics in Resource Management and Sustainable Ecosystem—Second International Conference, GRMSE 2014, Ypsilanti, MI, USA, 3–5 October 2014*; Bian, F., Xie, Y., Eds.; Springer: Berlin/Heidelberg, Germany, 2014; Volume 482, pp. 68–78. [CrossRef]
4. Liu, J.; Zhang, Y.; Liu, C.; Liu, X. Monitoring Impervious Surface Area Dynamics in Urban Areas Using Sentinel-2 Data and Improved Deeplabv3+ Model: A Case Study of Jinan City, China. *Remote Sens.* **2023**, *15*, 1976. [CrossRef]
5. Gan, Z.; Guo, S.; Chen, C.; Zheng, H.; Hu, Y.; Su, H.; Wu, W. Tracking the 2D/3D Morphological Changes of Tidal Flats Using Time Series Remote Sensing Data in Northern China. *Remote Sens.* **2024**, *16*, 886. [CrossRef]
6. Derakhshan, S.; Cutter, S.L.; Wang, C. Remote Sensing Derived Indices for Tracking Urban Land Surface Change in Case of Earthquake Recovery. *Remote Sens.* **2020**, *12*, 895. [CrossRef]
7. He, S.; Hua, M.; Zhang, Y.; Du, X.; Zhang, F. Forward Modeling of Scattering Centers From Coated Target on Rough Ground for Remote Sensing Target Recognition Applications. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 2000617. [CrossRef]
8. Han, L.; Paoletti, M.E.; Tao, X.; Wu, Z.; Haut, J.M.; Li, P.; Pastor, R.; Plaza, A. Hash-Based Remote Sensing Image Retrieval. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4411123. [CrossRef]
9. Hackstein, J.; Sumbul, G.; Clasen, K.N.; Demir, B. Exploring Masked Autoencoders for Sensor-Agnostic Image Retrieval in Remote Sensing. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5602914. [CrossRef]
10. Xu, S.; Ke, Q.; Peng, J.; Cao, X.; Zhao, Z. Pan-Denoising: Guided Hyperspectral Image Denoising via Weighted Represent Coefficient Total Variation. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5528714. [CrossRef]
11. Xu, S.; Cao, X.; Peng, J.; Ke, Q.; Ma, C.; Meng, D. Hyperspectral Image Denoising by Asymmetric Noise Modeling. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5545214. [CrossRef]
12. Xu, S.; Yu, C.; Peng, J.; Chen, S.; Cao, X.; Meng, D. Haar Nuclear Norms with Applications to Remote Sensing Imagery Restoration. *IEEE Trans. Image Process.* **2025**, *34*, 6879–6894. [CrossRef]
13. Xu, S.; Zhao, Z.; Bai, H.; Yu, C.; Peng, J.; Cao, X.; Meng, D. Hipandas: Hyperspectral image joint denoising and super-resolution by image fusion with the panchromatic image. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Honolulu, HI, USA, 19–24 October 2025; pp. 12002–12011.
14. Lin, J.; Huang, T.; Zhao, X.; Chen, Y.; Zhang, Q.; Yuan, Q. Robust Thick Cloud Removal for Multitemporal Remote Sensing Images Using Coupled Tensor Factorization. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5406916. [CrossRef]
15. Chen, Y.; Weng, Q.; Tang, L.; Zhang, X.; Bilal, M.; Li, Q. Thick Clouds Removing From Multitemporal Landsat Images Using Spatiotemporal Neural Networks. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4400214. [CrossRef]
16. Chen, Y.; Tang, L.; Yang, X.; Fan, R.; Bilal, M.; Li, Q. Thick Clouds Removal From Multitemporal ZY-3 Satellite Images Using Deep Learning. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 143–153. [CrossRef]
17. Li, L.; Huang, T.; Zheng, Y.; Zheng, W.; Lin, J.; Wu, G.; Zhao, X. Thick Cloud Removal for Multitemporal Remote Sensing Images: When Tensor Ring Decomposition Meets Gradient Domain Fidelity. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5512414. [CrossRef]
18. Peng, H.; Huang, T.; Zhao, X.; Lin, J.; Wu, W.; Li, L. Deep Domain Fidelity and Low-Rank Tensor Ring Regularization for Thick Cloud Removal of Multitemporal Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5409314. [CrossRef]
19. 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]
20. Zheng, W.J.; Zhao, X.L.; Zheng, Y.B.; Lin, J.; Zhuang, L.; Huang, T.Z. Spatial-spectral-temporal connective tensor network decomposition for thick cloud removal. *ISPRS J. Photogramm. Remote Sens.* **2023**, *199*, 182–194. [CrossRef]
21. Xu, S.; Wang, J.; Wang, J. Fast Thick Cloud Removal for Multi-Temporal Remote Sensing Imagery via Representation Coefficient Total Variation. *Remote Sens.* **2024**, *16*, 152. [CrossRef]
22. Xu, S.; Peng, J.; Ji, T.; Cao, X.; Sun, K.; Fei, R.; Meng, D. Stacked Tucker Decomposition with Multi-Nonlinear Products for Remote Sensing Imagery Inpainting. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5533413. [CrossRef]
23. Xu, S.; Zhao, Z.; Cao, X.; Peng, J.; Zhao, X.; Meng, D.; Zhang, Y.; Timofte, R.; Gool, L.V. Parameterized Low-Rank Regularizer for High-Dimensional Visual Data. *Int. J. Comput. Vis.* **2025**, *133* , 8546–8569. [CrossRef]
24. Sarukkai, V.; Jain, A.; Uzkent, B.; Ermon, S. Cloud Removal in Satellite Images Using Spatiotemporal Generative Networks. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision, WACV 2020, Snowmass, CO, USA, 1–5 March 2020; pp. 1785–1794. [CrossRef]
25. Hao, Y.; Jiang, W.; Liu, W.; Li, Y.; Liu, B. Selecting Information Fusion Generative Adversarial Network for Remote-Sensing Image Cloud Removal. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 6007605. [CrossRef]

26. Zhou, H.; Wang, Y.; Liu, W.; Tao, D.; Ma, W.; Liu, B. MSC-GAN: A Multistream Complementary Generative Adversarial Network with Grouping Learning for Multitemporal Cloud Removal. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5612014. [CrossRef]
27. Zou, X.; Li, K.; Xing, J.; Zhang, Y.; Wang, S.; Jin, L.; Tao, P. DiffCR: A Fast Conditional Diffusion Framework for Cloud Removal From Optical Satellite Images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5612014. [CrossRef]
28. Zhao, X.; Jia, K. Cloud Removal in Remote Sensing Using Sequential-Based Diffusion Models. *Remote Sens.* **2023**, *15*, 2861. [CrossRef]
29. Jing, R.; Duan, F.; Lu, F.; Zhang, M.; Zhao, W. Denoising Diffusion Probabilistic Feature-Based Network for Cloud Removal in Sentinel-2 Imagery. *Remote Sens.* **2023**, *15*, 2217. [CrossRef]
30. Long, C.; Yang, J.; Guan, X.; Li, X. Thick Cloud Removal from Remote Sensing Images Using Double Shift Networks. In Proceedings of the 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, Brussels, Belgium, 11–16 July 2021; pp. 2687–2690. [CrossRef]
31. Ebel, P.; Garnot, V.S.F.; Schmitt, M.; Wegner, J.D.; Zhu, X.X. UnCRtainTS: Uncertainty Quantification for Cloud Removal in Optical Satellite Time Series. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023—Workshops, Vancouver, BC, Canada, 17–24 June 2023; pp. 2086–2096. [CrossRef]
32. Zi, Y.; Song, X.; Xie, F.; Jiang, Z. Thick Cloud Removal in Multitemporal Remote Sensing Images Using a Coarse-to-Fine Framework. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 6005605. [CrossRef]
33. Long, C.; Li, X.; Jing, Y.; Shen, H. Bishift Networks for Thick Cloud Removal with Multitemporal Remote Sensing Images. *Int. J. Intell. Syst.* **2023**, *2023*, 9953198. [CrossRef]
34. Liu, H.; Huang, B.; Cai, J. Thick Cloud Removal Under Land Cover Changes Using Multisource Satellite Imagery and a Spatiotemporal Attention Network. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5601218. [CrossRef]
35. Li, Y.; Huang, Q.; Pei, X.; Jiao, L.; Shang, R. RADet: Refine Feature Pyramid Network and Multi-Layer Attention Network for Arbitrary-Oriented Object Detection of Remote Sensing Images. *Remote Sens.* **2020**, *12*, 389. [CrossRef]
36. Peng, D.; Zhang, Y.; Guan, H. End-to-End Change Detection for High Resolution Satellite Images Using Improved UNet++. *Remote Sens.* **2019**, *11*, 1382. [CrossRef]
37. He, N.; Fang, L.; Li, S.; Plaza, J.; Plaza, A. Skip-Connected Covariance Network for Remote Sensing Scene Classification. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *31*, 1461–1474. [CrossRef] [PubMed]
38. Han, Q.; Zhi, X.; Hu, J.; Zhang, S.; Chen, W.; Huang, Y.; Jiang, S. StyleFormer: Spatial-Temporal Style Projecting Bidirectional Interactive Transformer for Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5609016. [CrossRef]
39. Tang, X.; Li, M.; Ma, J.; Zhang, X.; Liu, F.; Jiao, L. EMTCAL: Efficient Multiscale Transformer and Cross-Level Attention Learning for Remote Sensing Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5626915. [CrossRef]
40. Huang, G.; Wu, P. CTGAN: Cloud Transformer Generative Adversarial Network. In Proceedings of the 2022 IEEE International Conference on Image Processing, ICIP 2022, Bordeaux, France, 16–19 October 2022; pp. 511–515. [CrossRef]
41. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef]
42. 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]
43. Ebel, P.; Xu, Y.; Schmitt, M.; Zhu, X. SEN12MS-CR-TS: A Remote Sensing Data Set for Multi-modal Multi-temporal Cloud Removal. *arXiv* **2022**, arXiv:2201.09613.
44. Zou, X.; Li, K.; Xing, J.; Tao, P.; Cui, Y. PMAA: A Progressive Multi-Scale Attention Autoencoder Model for High-Performance Cloud Removal from Multi-Temporal Satellite Imagery. In *Proceedings of the European Conference on Artificial Intelligence*; IOS Press: Amsterdam, The Netherlands, 2023; Volume 372, pp. 3165–3172. [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.