SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

March 2026, Vol. 69, Iss. 3, 132306:1–132306:15
https://doi.org/10.1007/s11432-025-4669-x

Change-aware multi-temporal cloud removal

Guochu YOU[1], Fang XU[2*], Jun PAN[3], Runmin DONG[4], Wen YANG[2,5] & Gui-Song XIA[1,2*]

[1]*School of Computer Science, Wuhan University, Wuhan 430072, China*
[2]*School of Artificial Intelligence, Wuhan University, Wuhan 430072, China*
[3]*State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, Wuhan 430072, China*
[4]*School of Artificial Intelligence, Sun Yat-sen University, Zhuhai 519082, China*
[5]*School of Electronic Information, Wuhan University, Wuhan 430072, China*

Received 17 March 2025/Revised 15 September 2025/Accepted 5 November 2025/Published online 20 January 2026

Abstract Cloud coverage varies significantly across different acquisitions, enabling the temporal complementarity of observations to recover surface details obscured by clouds. However, existing multi-temporal cloud removal methods fail to account for inconsistencies in data collected at different times (e.g., seasonal variations, human activities, meteorological events, etc.), resulting in severe performance degradation when surface changes occur. In this paper, we propose a change-aware multi-temporal cloud removal method, CA-MTCR, which utilizes SAR images, capable of penetrating clouds to capture terrain information beneath, to detect changes and optimize the usage of multi-temporal data. Specifically, we measure the similarity between SAR features to derive the change information for each auxiliary time step relative to the target time step, and then integrate this information into a spatial attention mechanism that highlights the regions requiring emphasis, thereby refining the multi-temporal fusion process. Furthermore, we propose a region-selective optical encoder to aggregate non-local information exclusively from partial cloud-free regions, enhancing the fidelity of the reconstructed images. Extensive evaluation demonstrates that the proposed algorithm outperforms state-of-the-art cloud removal algorithms with a gain of about 2.0 dB in terms of PSNR on the SEN12MS-CR-TS dataset and 0.8 dB on the Allclear dataset, and exhibits robustness to variations in the length of auxiliary time steps and their offset relative to the target time step.

Keywords remote sensing, cloud removal, optical and SAR images, multi-temporal, multi-modal

Citation You G C, Xu F, Pan J, et al. Change-aware multi-temporal cloud removal. Sci China Inf Sci, 2026, 69(3): 132306, https://doi.org/10.1007/s11432-025-4669-x

1 Introduction

Optical remote sensing employs visible, near-infrared, and shortwave-infrared sensors to capture images of the Earth's surface, and continues to be the most extensively utilized technique in the field of Earth observation [1–6]. However, optical remote sensing images inevitably suffer from cloud interference, which substantially diminishes their capacity to accurately represent ground object information [7–9]. Cloud removal, which aims to reconstruct cloud-obscured information, is critical for the stable application of optical remote sensing technology across diverse fields [10–14]. Due to the erasure of textures in cloud-covered regions, cloud removal is an inherently ill-posed problem that typically requires the integration of auxiliary reference data to achieve reliable results [15–18].

Multi-temporal optical images, obtained through periodic satellite revisits, provide a promising approach to compensating for cloud-obscured regions, because of the dynamic movement and morphological changes of clouds across different time phases. Multi-temporal cloud removal has been extensively studied and made great progress in the past few years [19–22]. However, the majority of previous methods primarily treat multi-temporal images uniformly, indiscriminately aggregating all cloud-free regions to generate a composite cloud-free image [23–25], as shown in Figure 1(a). This approach often overlooks the potential discrepancies between different time steps, leading to artifacts or misaligned features in the final reconstruction result. In contrast, SAR (synthetic aperture radar), with the characteristic of penetrating clouds, can capture information about the terrain beneath the clouds [26–28]. By utilizing multi-temporal SAR images, as shown in Figure 1(b), it is feasible to accurately identify which regions undergo changes and which remain relatively stable, facilitating the targeted selection of data from stable regions for cloud-free image reconstruction.

* Corresponding author (email: xufang@whu.edu.cn, guisong.xia@whu.edu.cn)

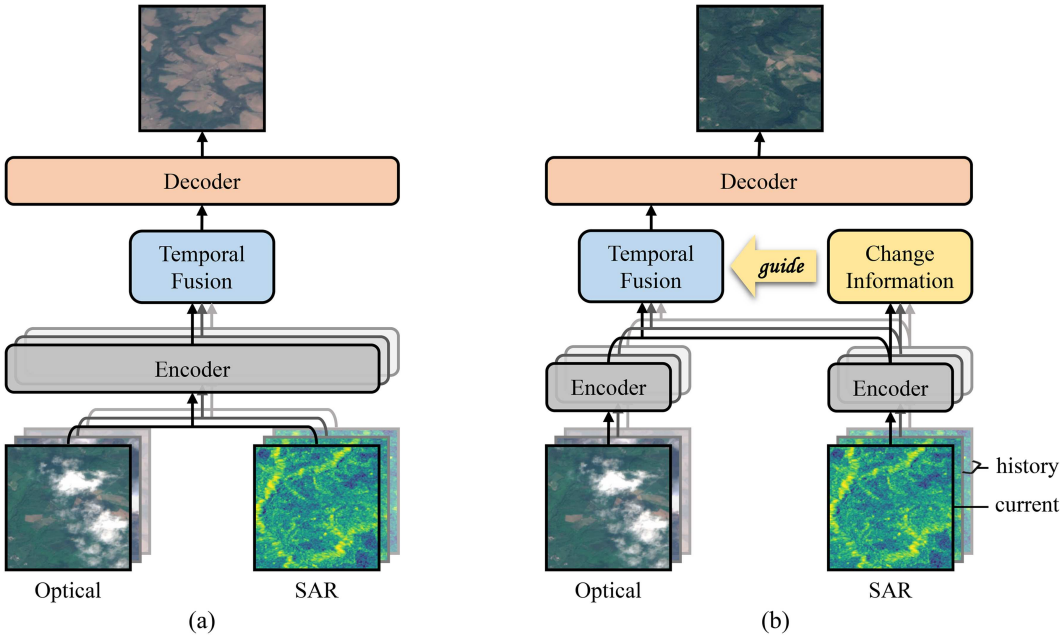

Figure 1 (Color online) Comparisons of different learning paradigms for multi-temporal cloud removal. (a) UnCRtainTS treats multi-temporal images uniformly, neglecting the inherent inconsistencies in data collected at different times, which results in artifacts or misaligned features in the final reconstruction results; (b) our proposed CA-MTCR utilizes SAR images to detect changes in ground objects, thereby optimizing the use of multi-temporal data and boosting the quality of the reconstruction.

In this paper, we propose a novel change-aware multi-temporal cloud removal algorithm, termed CA-MTCR, to optimize the utilization of multi-temporal data by leveraging changes in ground objects detected from SAR images. Specifically, we begin by maximizing the extraction of valuable information from both optical and SAR images at each temporal phase. For optical images, we propose a region-selective encoder that focuses on aggregating non-local information from cloud-free areas, effectively minimizing interference from cloud-covered regions during feature extraction. For SAR images, we implement an encoder pretrained using the Noise2Noise technique [29], tailored to alleviate the information blurring caused to speckle noise. To enhance the informational richness of each temporal phase, the optical and SAR features are then integrated into a unified representation. Subsequently, we propose a change-aware multi-temporal fusion module to integrate features with temporal coherence. By measuring feature similarity within SAR data, we capture dynamic changes in the auxiliary multi-temporal data relative to the target time step, which serve to guide the integration of valuable information across temporal phases and generate a comprehensive feature representation for high-quality cloud-free image reconstruction.

To sum up, the contributions of this work are threefold.

• We propose a novel change-aware multi-temporal cloud removal method, CA-MTCR, which leverages SAR data to capture changes over time, optimizing the utilization of multi-temporal data and remarkably enhancing cloud-free image reconstruction performance.

• We propose a region-selective optical encoder that concentrates on valuable information from cloud-free regions while mitigating interference introduced by cloudy areas, based on which CA-MTCR effectively extracts and integrates the detailed features of cloud-free regions.

• Extensive experiments show that the proposed method achieves state-of-the-art results on the SEN12MS-CR-TS and AllClear datasets, exhibiting robustness to variations in the length of auxiliary time steps and their offset relative to the target time step.

2 Related work

Cloud removal, a long-standing research problem, aims to reconstruct the missing information caused by clouds in optical satellite imagery. Early attempts often assume that cloud-covered and cloud-free regions exhibit similar statistical and geometric characteristics [30,31], i.e., spatial-based methods, or utilize spectral bands with relatively strong penetration capabilities through clouds as auxiliary information [32–36], i.e., spectral-based methods. However, these methods rely on limited known information, resulting in significant uncertainty in reconstruction results. Moreover, as cloud thickness increases, all the land signals in the optical bands are obstructed. Overall, relying

solely on a single cloud-obscured optical image itself, due to its limited valuable spatial and spectral information, results in constrained reconstruction capabilities, particularly in areas with extensive thick cloud cover or complex, heterogeneous land cover types. To overcome these limitations, SAR images are often integrated, as they provide critical information that optical bands cannot capture under cloud cover [9,11,37–39]. For example, DSen2-CR [39] concatenates the SAR image to the input optical image and uses a deep residual neural network to predict the target cloud-free optical image. Since SAR and optical images reveal different characteristics of observed objects—most notably, SAR images lack spectral information—it is challenging to transform SAR images into spectral and texture features similar to those of optical images, resulting in suboptimal spectral fidelity in the reconstruction results.

Remote sensing satellites feature periodic revisit capabilities, and the movement and morphological transformations of clouds allow data from different time phases to potentially provide clear surface information for regions obscured by cloud cover. Consequently, a few studies have explored the means of multi-temporal data fusion, integrating information across time to facilitate cloud removal [40–44]. For example, Gao et al. [45] utilized tempo-spectral angle mapping (TSAM) to measure the similarity between pixels across spectral and temporal dimensions, and applied a multi-temporal replacement method to recover missing data with the pixels selected by TSAM. It heavily depends on valid historical optical observations and assumes temporal consistency, which renders it unreliable in regions characterized by frequent cloud cover or by rapid land cover changes. Rather than relying on pixel replacement, STGAN [23] concatenates multi-temporal cloudy images as input to a spatiotemporal generator network to generate a cloud-free image. Czerkawski et al. [46] employed historical cloud-free optical data as prior information to reconstruct regions obscured by clouds. To enhance the focus on cloud-free regions within multi-temporal images for cloud-free image reconstruction, UnCRtainTS [25] employs attention-based temporal aggregation to combine the sequence of observations. However, these methods overlook the potential inconsistencies in data collected at different times, which may arise from changes in ground objects caused by seasonal variations or human activities. As a result, they struggle to effectively address the degradation in reconstruction performance caused by surface changes. In this paper, we utilize SAR images to detect changes in ground objects, thereby optimizing the use of multi-temporal data. It is worth noting that the role of SAR in our framework differs fundamentally from that in mono-temporal SAR-based cloud removal approaches [11,15,39], where the reconstruction of cloud-obscured regions primarily relies on transforming the corresponding regions in SAR imagery into spectral and textural representations similar to those of optical images, resulting in degraded spectral fidelity. In contrast, since our work leverages multi-temporal SAR data to optimize the use of multi-temporal observations, historical optical imagery provides reliable cues for reconstruction in unchanged regions, whereas only regions undergoing changes rely on SAR imagery to approximate spectral and textural attributes, thereby minimizing spectral fidelity distortions.

In parallel with efforts to leverage diverse data sources, another line of work explores advanced network architectures to improve cloud removal performance. Recent advances follow convolutional neural network (CNN), and numerous CNN-based models have been proposed to enhance reconstruction quality [24, 25, 39]. For instance, DSen2-CR [39] is derived from the EDSR super-resolution network and integrates SAR information to predict cloud-free optical imagery. UnCRtainTS [25] relies on efficient MBConv blocks [47], which combine depthwise convolutions with pointwise convolutions to enable computationally efficient spatial encoding, while aggregating temporal information across multiple observations. Since cloud-free regions in the optical image at the target time step provide critical cues for reconstruction, the limited receptive field of convolutional networks constrains their ability to integrate non-local contextual information. To mitigate this limitation, some studies integrate modules such as dilated convolutions [48,49] and multi-scale feature fusion [19] to explicitly enlarge the receptive field. More recently, Transformer-based or hybrid CNN-Transformer architectures [9, 11, 15, 22] are proposed, which leverage the Transformer's capability of modeling long-range dependencies to more effectively integrate non-local contextual information. For example, HDRSA-Net [9] is built on a hybrid architecture that integrates dynamic local exploration modules with sparse global context aggregation, enabling the extraction of fine-grained local features while maintaining the ability to capture long-range dependencies. While these methods exhibit strong feature extraction capabilities to facilitate the recovery of cloud-free imagery, they tend to introduce unnecessary redundant features, since cloud-covered regions contribute little meaningful information yet still propagate through the network, leading to the blurring or weakening of critical information. To overcome this issue, we propose a region-selective optical encoder within our framework that explicitly emphasizes informative cloud-free regions while suppressing the influence of cloud-contaminated areas, thereby strengthening the representation of critical features and ultimately enhancing the reliability and spectral fidelity of the reconstructed imagery.

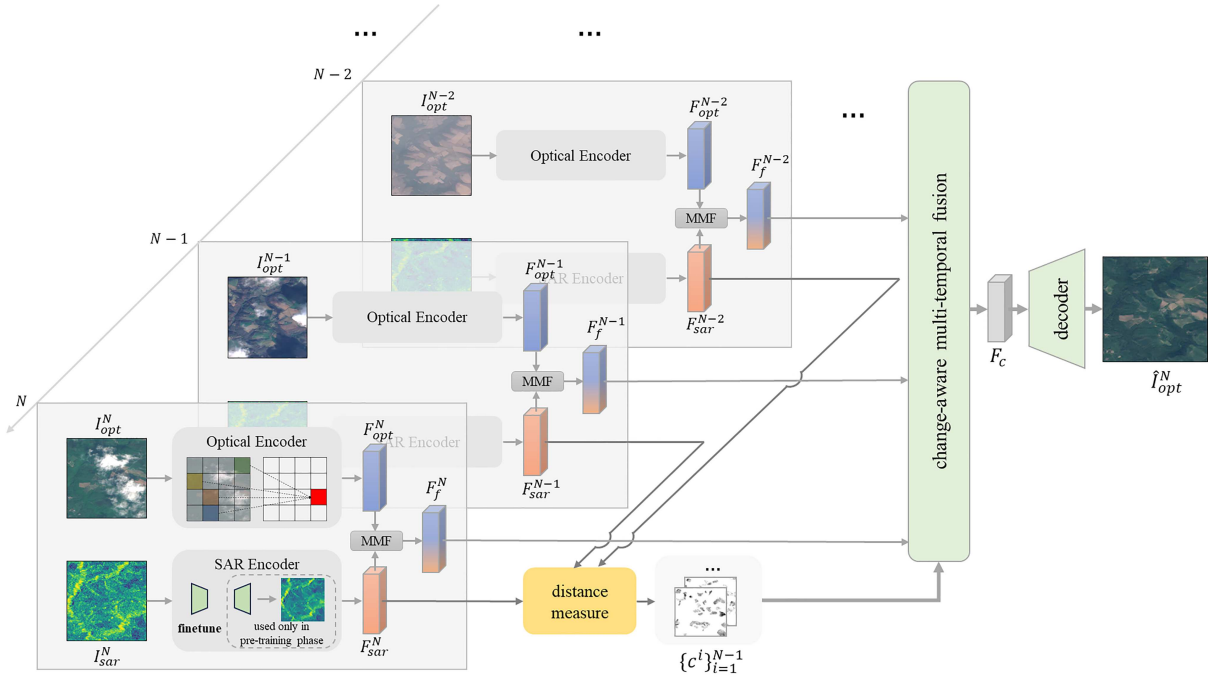

Figure 2 (Color online) Overview of the proposed CA-MTCR method.

3 CA-MTCR

3.1 Overview

Problem statement. Given a time series of optical images $\{I^i_{\text{opt}}\}_{i=1}^N$ and their corresponding SAR images $\{I^i_{\text{sar}}\}_{i=1}^N$, the task of cloud removal aims to reconstruct the clear optical image $\hat{I}^t_{\text{opt}}$ corresponding to the cloud-affected optical image I^t_{opt}, for $t \in 1, \ldots, N$. Specifically, in this paper, we focus on reconstructing the optical image at $t = N$, i.e., using current and historical data to reconstruct the optical image at the current time step. Existing multi-temporal cloud removal methods typically integrate all available data to generate a cloud-free image

$$\hat{I}^N_{\text{opt}} = \mathbf{F}_{\text{MTCR}}\left(\left\{\left(I^i_{\text{opt}}, I^i_{\text{sar}}\right)\right\}_{i=1}^N\right), \tag{1}$$

where $\mathbf{F}_{\text{MTCR}}$ is a multi-temporal cloud removal model that takes the entire time series of optical and SAR images as a unified input, without accounting for the variations between different time steps. However, due to factors such as climate change, human activities, and seasonal variations, the ground information may change over time, leading to the generated cloud-free optical image failing to accurately reflect the true content of the scene at $t = N$. In contrast, we distinguish between the current time step and historical time steps

$$\hat{I}^N_{\text{opt}} = \mathbf{F}_{\text{CA-MTCR}}\left(\left(I^N_{\text{opt}}, I^N_{\text{sar}}\right) \mid \left\{\left(I^i_{\text{opt}}, I^i_{\text{sar}}\right)\right\}_{i=1}^{N-1}\right), \tag{2}$$

where $\mathbf{F}_{\text{CA-MTCR}}$ is a change-aware multi-temporal cloud removal model that explicitly identifies the target time step to reconstruct, while using data from other time steps as references rather than treating them equally. By accounting for the inherent differences between the current and historical time steps, we are able to capture the temporal dynamics and leverage consistent historical data, filtering out inconsistent information to enhance cloud removal performance.

Architecture. The overall architecture of the proposed CA-MTCR algorithm is illustrated in Figure 2. It begins by processing each time step independently. Specifically, for each time step t, it extracts features from the optical and SAR images using two tailored encoders, $\mathbf{F}_{\text{opt}}(\cdot)$ and $\mathbf{F}_{\text{sar}}(\cdot)$, respectively

$$F^t_{\text{opt}} = \mathbf{F}_{\text{opt}}\left(I^t_{\text{opt}}\right), t \in \{1, \ldots, N\}, \tag{3}$$

$$F^t_{\text{sar}} = \mathbf{F}_{\text{sar}}\left(I^t_{\text{sar}}\right), t \in \{1, \ldots, N\}, \tag{4}$$

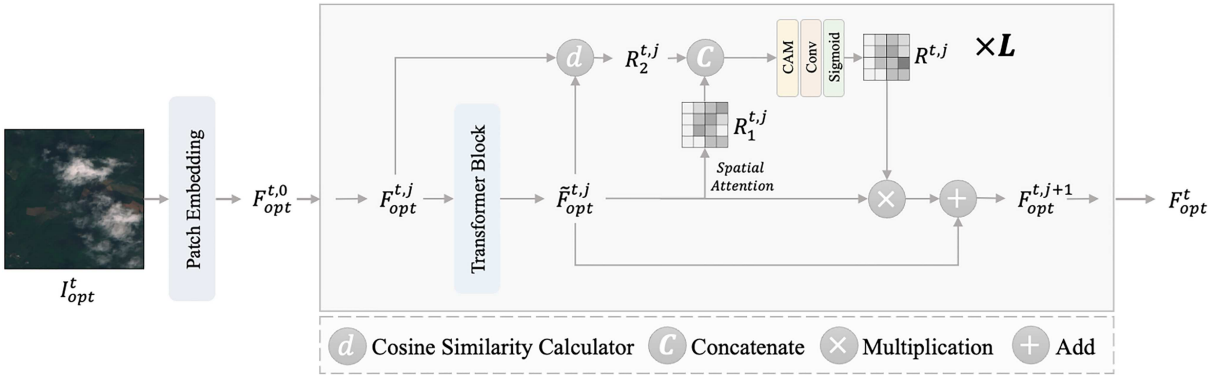

Figure 3 (Color online) Detail of the region-selective optical encoder.

where $\mathbf{F}_{\text{opt}}(\cdot)$ is designed to emphasize information from cloud-free regions while systematically disregarding that from cloud-covered areas, and $\mathbf{F}_{\text{sar}}(\cdot)$ is optimized to mitigate the impact of speckle noise, thereby extracting more reliable features. F^t_{opt} and F^t_{sar} are then integrated via a multi-modal fusion block, $\mathbf{H}_{\text{mmf}}(\cdot)$, to enhance the integrity of single-temporal information, as follows:

$$F^t_f = \mathbf{H}_{\text{mmf}}\left(F^t_{\text{opt}}, F^t_{\text{sar}}\right), t \in \{1, \ldots, N\}. \tag{5}$$

Before fusing the multi-temporal features $\{F^i_f\}^N_{i=1}$, the change information c^t for each historical time step t relative to the current time step N is computed based on the SAR features

$$c^t = \mathbf{H}_{\text{change}}\left(F^t_{\text{sar}}, F^N_{\text{sar}}\right), t \in \{1, \ldots, N-1\}, \tag{6}$$

where $\mathbf{H}_{\text{change}}(\cdot)$ is the function to capture the temporal differences between the historical and current time steps. Subsequently, the multi-temporal features $\{F^i_f\}^N_{i=1}$ are aggregated based on the change information $\{c^i\}^{N-1}_{i=1}$, using a change-aware multi-temporal fusion module $\mathbf{H}_{\text{CAF}}(\cdot)$, to generate a comprehensive feature representation, F_c, as follows:

$$F_c = \mathbf{H}_{\text{CAF}}\left(\left\{F^i_f\right\}^N_{i=1}, \left\{c^i\right\}^{N-1}_{i=1}\right). \tag{7}$$

Finally, the high-quality cloud-free image $\hat{I}^N_{\text{opt}}$ is reconstructed from the comprehensive feature representation F_c

$$\hat{I}^N_{\text{opt}} = \mathbf{H}_d\left(F_c\right), \tag{8}$$

where $\mathbf{H}_d$ represents the decoder function specifically designed for cloud-free image reconstruction.

3.2 Mono-temporal feature extraction and fusion

Accurate extraction of meaningful information from multi-modal images is a fundamental prerequisite for achieving high-quality cloud-free image reconstruction. In this work, we employ customized encoders designed to extract features specifically tailored to the distinctive characteristics of optical and SAR images.

Optical encoder. Cloud cover manifests in optical images as "information blind areas". By leveraging the spatial context information of cloud-free regions, it is possible to infer the underlying ground information obscured by the clouds. However, the influence of these blind areas can propagate across the entire image, potentially compromising the accurate representation of information within the unobscured regions. To maximize the extraction and utilization of valuable information from uncontaminated regions in optical images, we propose a region-selective optical encoder to aggregate non-local information exclusively from partially unobscured regions, as shown in Figure 3. It leverages the Transformer's ability to efficiently propagate information across the entire image to accumulate long-range varying contextual information. To ensure that only valuable information is aggregated, we incorporate a region-selection mechanism that focuses on relevant regions for each Transformer block. Let $F^{t,j}_{\text{opt}}$ and $\tilde{F}^{t,j}_{\text{opt}}$ denote, respectively, the input and output of the j-th Transformer block for the optical image I^t_{opt}. We distinguish these valuable regions in two ways.

• Since features corresponding to cloudy regions exhibit abnormal responses, we apply a spatial attention module on the output of the Transformer block, $\tilde{F}^{t,j}_{\text{opt}}$, to highlight these areas, with the attention weights denoted by $R^{t,j}_1$.

• Since the learning objective is to preserve cloud-free areas while filling in the cloudy parts, the difference between the input and output of the Transformer block, where significant changes occur, reveals the cloudy regions. Specifically, we measure these differences by computing the cosine distance between $F_{\text{opt}}^{t,j}$ and $\tilde{F}_{\text{opt}}^{t,j}$, and denote the result by $R_2^{t,j}$.

We then concatenate $R_1^{t,j}$ and $R_2^{t,j}$, and apply a channel attention module (CAM) to weight them. A convolution layer followed by a sigmoid activation is used to learn a soft mask $R^{t,j}$. Finally, we adopt the mask to constrain the output of the Transformer block $\tilde{F}_{\text{opt}}^{t,j}$ before passing it to the next block, suppressing features from the cloudy regions to avoid disturbing the information from the cloud-free regions. After passing through L Transformer blocks, the resulting optical feature $F_{\text{opt}}^{t,L}$ serve as the final optical feature F_{opt}^{t}.

SAR encoder. SAR images suffer severely from speckle noise [50], which appears grainy with dark and bright spots, making the representation of SAR features blurred and chaotic. Given that SAR features are indispensable for capturing temporal changes and thereby supporting the effective fusion of multi-temporal data, mitigating the adverse effects of speckle noise is of critical importance. To this end, we pretrain the SAR encoder with an SAR denoising network prior to its integration into the proposed framework. After training, only the encoder is retained to initialize the SAR feature extractor, while the decoder and other components are discarded. Since noise-free SAR ground truth is unavailable, we adopt the self-supervised Noise2Noise method [29] for pretraining, which learns to suppress noise by leveraging pairs of noisy images from the same scene. Specifically, we implement the denoising network using the VQ-VAE architecture [51], commonly employed for learning compact representations and removing high-frequency noise, and leverage SAR images from different temporal phases to learn noise-independent features. In this work, the SAR encoder is constituted by three convolution layers followed by two MBConv blocks [47]. The pretraining objective combines an $L2$ reconstruction loss with latent regularization terms, defined as

$$L_{\text{pre}} = \|y_2 - \hat{y_1}\|_2 + \alpha \left\|\text{sg}[z_e] - e_c\right\|_2^2 + \beta \left\|z_e - \text{sg}[e_c]\right\|_2^2, \tag{9}$$

where $\text{sg}[\cdot]$ denotes the stop-gradient operator. y_1 and y_2 are two SAR images of the same scene with different noise, and $\hat{y_1}$ denotes the reconstruction of y_1. z_e is the output of the SAR encoder and e_c is the learnable codebook of discrete embeddings in VQ-VAE.

Multi-modal fusion block. Due to the inherent uncertainty associated with cloud cover, the information provided by multi-temporal optical images may not adequately compensate for the information loss at time step N. Consequently, before proceeding with multi-temporal data fusion, we enhance the informational completeness of each time step by fusing optical feature with SAR feature. Specifically, we concatenate the optical and SAR features, and subsequently feed the concatenated features into a convolutional layer as well as an MBConv layer.

3.3 Change-aware multi-temporal fusion

The detail of the change-aware multi-temporal fusion module is shown in Figure 4. To counteract the degradation in reconstruction performance caused by land cover changes, we calculate the cosine similarity between SAR features at each historical time step t relative to the current time step N, i.e., c^t, to obtain change information. Specifically, the greater the cosine similarity, the lower the likelihood of significant changes between the time steps. For the current time step N, we also assign a corresponding c^t, which is a matrix of ones, to facilitate subsequent calculations

$$c^t = \begin{cases} \dfrac{\left\langle F_{\text{sar}}^t, F_{\text{sar}}^N \right\rangle}{\|F_{\text{sar}}^t\| \|F_{\text{sar}}^N\|}, & 1 \leqslant t \leqslant N-1, \\ \mathbf{1}, & t = N. \end{cases} \tag{10}$$

We then utilize this change information to guide the fusion of multi-temporal features. Specifically, for each time step t, we compute a corresponding spatial attention weight to highlight the most valuable information, i.e., that which is temporally consistent and cloud-free

$$\left\{W^t\right\}_{t=1}^N = \text{Attn}\left(\left\{\text{Conv}\left(c^t \parallel F_f^t\right)\right\}_{t=1}^N\right), \tag{11}$$

where $\parallel$ denotes the concatenation operation, Conv represents a convolution layer, and Attn refers to the function responsible for computing the attention weight. In this paper, we adopt the L-TAE [52], which has consistently served as a reliable foundation in recent studies for attention modeling across multi-temporal remote sensing imagery [53, 54], to conduct the weight computation process. It enables to capture of location-specific temporal

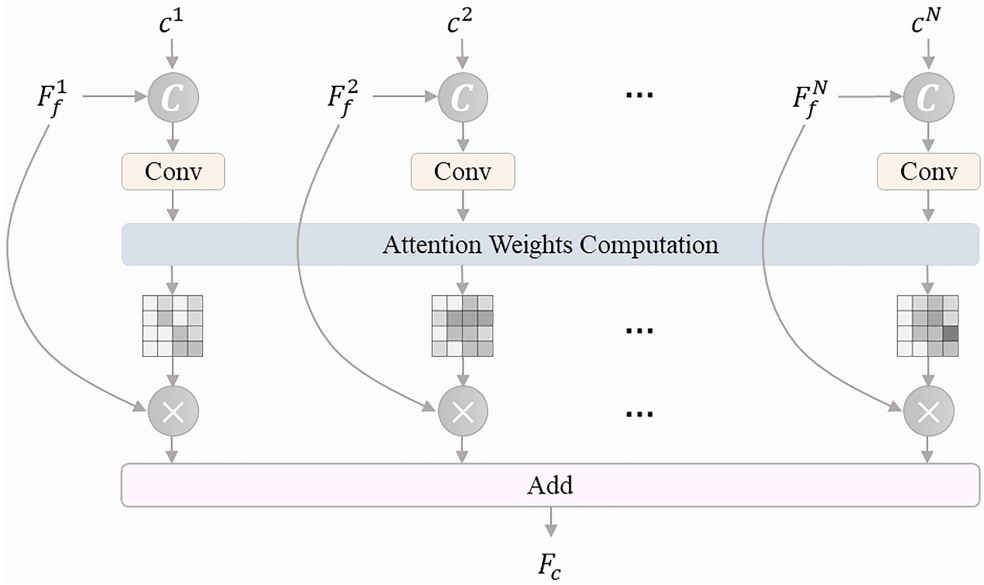

Figure 4 (Color online) Detail of the change-aware multi-temporal fusion module.

dynamics and is optimized for memory and computational efficiency. With the weights, we effectively merge the features from different time sequences to form a comprehensive feature representation

$$F_c = \sum_{t=1}^{N} \left(F_f^t \odot W^t \right), \tag{12}$$

where $\odot$ denotes the element-wise multiplication operation, achieved through broadcasting along the channel dimension.

3.4 Loss function

We train the CA-MTCR network by minimizing the discrepancy between its output, $\hat{I}_{\text{opt}}^N$, and the reference cloud-free image I_{opt}^{cf}, which is temporally close to the cloudy image at the time step N. We use the L1 loss as the primary optimization objective for cloud removal, denoted by L_{l1}, which is formulated as

$$L_{l1} = \left\| \hat{I}_{\text{opt}}^N - I_{\text{opt}}^{cf} \right\|_1. \tag{13}$$

Additionally, two types of perceptual losses are incorporated. (i) Feature reconstruction loss, denoted as L_{fr}, encourages the restoration of sharp edges and fine details, thereby improving the structural integrity and visual quality of the reconstructed cloud-free images. Specifically, L_{fr} is computed using the Manhattan distance between feature representations extracted by the VGG16 network from both the reconstructed and the reference cloud-free images. (ii) Style reconstruction loss, denoted as L_{sr}, encourages fidelity in the spectral and texture properties of the reconstructed image. Specifically, L_{sr} is computed using the squared Frobenius norm of the difference between the Gram matrices of the output and target cloud-free images. The corresponding formulations are as follows:

$$L_{\text{fr}} = \sum_{l=1}^{L} \left\| \Phi_l \left(\hat{I}_{\text{opt}}^N \right) - \Phi_l \left(I_{\text{opt}}^{cf} \right) \right\|_1, \tag{14}$$

$$L_{\text{sr}} = \sum_{l=4}^{L} \left\| G \left(\Phi_l \left(\hat{I}_{\text{opt}}^N \right) \right) - G \left(\Phi_l \left(I_{\text{opt}}^{cf} \right) \right) \right\|_1, \tag{15}$$

where $\Phi_l(\cdot)$ refers to the feature extraction operation associated with the l-th layer of the pretrained VGG16 network, $G(\cdot)$ denotes the computation of the Gram matrix that captures feature correlations for style representation, and L is the total number of selected layers. The overall function is as follows:

$$L = \lambda_1 L_{l1} + \lambda_2 L_{\text{fr}} + \lambda_3 L_{\text{sr}}, \tag{16}$$

where λ_1, λ_2 and λ_3 denote the balancing parameters.

4 Experiments

4.1 Experimental settings

Dataset and metrics. The experiments are conducted on SEN12MS-CR-TS [24] and AllClear [21], two large-scale benchmark datasets that provide multi-modal and multi-temporal observations specifically curated for cloud removal.

SEN12MS-CR-TS. It comprises Sentinel-1 SAR and Sentinel-2 optical observations collected from 53 globally distributed regions of interest (ROIs), with 35 ROIs allocated for training, 5 ROIs for validation, and 13 ROIs for testing. By means of a non-overlapping sliding window of 256×256 pixels, 10176 training samples, 1410 validation samples, and 3716 test samples are derived, each containing 30 time steps of coregistered optical and SAR images. We select cloudy and cloud-free image pairs with minimal land cover changes, where the time step of the cloudy image is assigned as the one to be restored, and the cloud-free image serves as its corresponding target. Additional time steps are randomly selected from the remaining time steps to serve as auxiliary data. For preprocessing, the optical data are clipped to the range $[0, 10000]$ and scaled to $[0, 1]$, and the SAR data are clipped to the range $[-25, 0]$ and scaled to $[0, 1]$ as well.

AllClear. It comprises 23742 globally distributed ROIs, each containing multi-spectral optical imagery from Sentinel-2 and SAR imagery from Sentinel-1. In this paper, we adopt AllClear (3.4%) as the training set, which is matched in size to the training set of the SEN12MS-CR-TS dataset, including 10176 training samples and 1500 validation samples. For evaluation, the test set consists of 2735 samples. Each sample provides data across four consecutive time steps, one of which contains a cloud-free optical image serving as the ground truth (GT). In other words, the construction of the AllClear dataset explicitly addresses the temporal misalignment problem by ensuring that the auxiliary time steps are selected to be as close as possible to the target, thereby minimizing potential land cover changes between the auxiliary inputs and the target observation. It differs from SEN12MS-CR-TS, where only the GT is chosen adjacent to the target time step to avoid evaluation bias, but no constraints are placed on the auxiliary observations, making the dataset more aligned with practical use. All preprocessing steps are kept consistent with those of the SEN12MS-CR-TS dataset.

The performance of cloud removal is evaluated using four metrics: peak signal-to-noise ratio (PSNR) in decibels (dB), spectral angle mapper (SAM) in degrees (°), structural similarity index measure (SSIM) which is unitless, and mean absolute error (MAE) in top-of-atmosphere reflectance (ρ_{TOA}).

Implementation details. The proposed network, CA-MTCR, is implemented in Pytorch and trained on two NVIDIA 3090 GPUs. The batch size is set to 6 and the maximum epoch of training iterations is set to 30. The Adam optimizer is employed with a learning rate of 1×10^{-4} for the entire network, except for the pretrained SAR encoder, which is fine-tuned at a smaller learning rate of 1×10^{-5}. The learning rates are scheduled to decay by 50% every 5 epochs following the initial 10 epochs. The weighting factors λ_1, λ_2 and λ_3 are empirically set to 2.0, 1.0 and 250.0. Additionally, the number of Transformer blocks in the optical encoder L is configured to 9. The perceptual loss is computed using VGG16 features from `relu11`, `relu21`, `relu31`, `relu41`, and `relu51`, i.e., L within L_{fr} and L_{sr} is set to 5. For the pretraining of the SAR encoder, SAR pairs that correspond to selected cloudy and cloud-free optical image pairs with minimal temporal variation are used. The encoder is optimized for 50 epochs with a learning rate of 2×10^{-4}, with the regularization weights α and β within L_{pre} set to 1 and 0.1, respectively.

4.2 Comparison with state-of-the-arts

We compare the proposed CA-MTCR network to state-of-the-art cloud removal methods, including mono-temporal approaches, DSen2-CR [39], Align-CR [11] and HDRSA-Net [9], and multi-temporal approaches, STGAN [23], U-TAE [55], CR-TS-Net [24] and UnCRtainTS [25]. To ensure a fair comparison, STGAN [23], originally designed for multi-temporal optical images, is adapted to incorporate multi-temporal optical and SAR image pairs. Additionally, U-TAE [55], a state-of-the-art remote sensing image time series encoder, is included as a baseline with minor modifications tailored for cloud removal. For the multi-temporal approaches, we randomly select two time steps as auxiliary inputs to assist in the restoration of the target time step.

Quantitative results. The results are shown in Table 1. Our proposed CA-MTCR network demonstrates significant performance gains over existing state-of-the-art methods. Specifically, on the SEN12MS-CR-TS dataset, mono-temporal approaches that incorporate SAR data as auxiliary information demonstrate promising results, as SAR data alone can effectively compensate for information gaps in optical images caused by cloud coverage. In contrast, existing multi-temporal approaches, despite leveraging data from additional time steps, fail to surpass the performance of mono-temporal approaches. This limitation stems from their uniform treatment of multi-temporal

Table 1 Quantitative comparisons of the proposed CA-MTCR network to state-of-the-art methods on the SEN12MS-CR-TS and AllClear datasets, where PSNR is measured in decibels (dB), SAM in degrees, SSIM is unitless, MAE in reflectance units (ρ_{TOA}), and model complexity is reported by parameters (M) and FLOPs (G). The best-performing results are highlighted in bold.

Method	Type	SEN12MS-CR-TS				AllClear				Params	FLOPs
		PSNR↑	SAM↓	SSIM↑	MAE↓	PSNR↑	SAM↓	SSIM↑	MAE↓		
DSen2-CR [39]	Mono-temporal	29.47	6.27	0.878	0.025	30.66	7.75	0.875	0.028	18.95	1241.18
Align-CR [11]		30.53	5.76	0.895	0.022	30.95	7.92	0.876	0.028	45.56	450.63
HDRSA-Net [9]		30.76	5.60	0.903	0.022	31.00	7.74	0.888	0.028	21.43	2806.60
STGAN [23]	Multi-temporal	26.27	10.84	0.831	0.036	31.31	7.24	0.869	0.024	93.53	15.41
U-TAE [55]		29.34	7.31	0.892	0.026	32.45	6.73	0.897	0.022	1.03	14.38
CR-TS Net [24]		29.68	6.08	0.894	0.025	33.30	5.13	0.920	0.020	38.47	7550.35
UnCRtainTS [25]		29.76	6.75	0.900	0.024	33.41	5.49	**0.925**	0.021	0.57	37.85
CA-MTCR (ours)		**31.79**	**5.18**	**0.903**	**0.019**	**34.21**	**4.85**	0.921	**0.018**	7.45	456.57

images, where the noise introduced by extra temporal data—such as changes in land cover and cloud dynamics—outweighs the benefits of additional information, resulting in diminished overall effectiveness. Our method addresses this issue by utilizing SAR imagery to detect changes and optimize the use of multi-temporal data, effectively enhancing performance beyond that of mono-temporal cloud removal approaches. Moreover, we observe greater improvements in metrics such as PSNR, SAM, and MAE compared to SSIM. This discrepancy arises because SAR data lack spectral information, which, while effective in compensating for the structural aspects of optical images like land contours and textures, struggle to restore the unique spectral attributes of optical images, such as color and reflectance. Overall, CA-MTCR can fully exploit the valuable information from multi-temporal data and resists interference from the additional noise introduced by multi-temporal data, achieving the best results in terms of spectral fidelity and structural integrity.

On the AllClear dataset, where temporal misalignment issues are explicitly mitigated to minimize unintended variations, multi-temporal approaches generally outperform mono-temporal ones, with our method achieving the best performance. In addition, we can observe that the performance improvement of our method over the best-performing baseline is more pronounced on the SEN12MS-CR-TS dataset (2.03 dB in PSNR) than in the AllClear dataset (0.80 dB in PSNR). This discrepancy is primarily attributed to the fact that our method accounts for inconsistencies in data collected at different times, which further highlights its practicality, as it does not impose strict requirements on the quality or alignment of auxiliary data and remains effective in challenging real-world scenarios. Meanwhile, our method remains relatively compact in parameter size and moderate in computational complexity, without introducing excessive overhead.

Qualitative results. We showcase two typical scenes from each of the SEN12MS-CR-TS and AllClear datasets for qualitative analysis, as illustrated in Figures 5 and 6, respectively. To better illustrate differences in reconstruction quality across methods, for each scene, we highlight two specific regions with enlarged views to enhance the visualization of fine structural details, and include PSNR heatmaps for each method. We can find that our proposed method can effectively restore surface information obscured by cloud cover. The resulting images exhibit high integrity in land contours, textures, and spatial structures, and demonstrate exceptional spectral fidelity. Mono-temporal methods, due to the lack of adequate auxiliary information for compensation, tend to generate blurring in areas covered by clouds, resulting in the loss or distortion of terrain details. For instance, as illustrated in the first scene of Figure 5 where thick cloud cover obscures most of the area, DSen2-CR and Align-CR fail to adequately reconstruct intricate landscape features, particularly the delicate curvilinear patterns, since they rely solely on the very limited information available in the optical image and the corresponding SAR image for restoration. Existing multi-temporal approaches, with access to data from additional time steps, are capable of reconstructing more complete terrain details. However, these methods are susceptible to inconsistencies and dynamic cloud interferences arising from temporal variations across different time steps, leading to the generation of numerous artifacts. For example, as demonstrated in the second scene of Figure 5, existing multi-temporal methods tend to utilize the optical images from auxiliary time step 2, which are free from cloud contamination, to reconstruct the cloudy optical image. Nevertheless, due to seasonal variations, the auxiliary optical image exhibits significant discrepancies in the appearance of agricultural land compared to the target time step. Consequently, the reconstructed image fails to accurately reflect the actual conditions at the target time. By leveraging SAR images, regions that have undergone changes can be clearly identified. Our proposed CA-MTCR utilizes the change information to guide the fusion of multi-temporal data, enabling robust reconstruction that is resistant to temporal inconsistencies. Moreover, we observe that in homogeneous regions, existing methods tend to produce speckled or mottled artifacts, as shown in Figure 6. This issue can be attributed to the speckle noise intrinsic to SAR imagery, which propagates into the

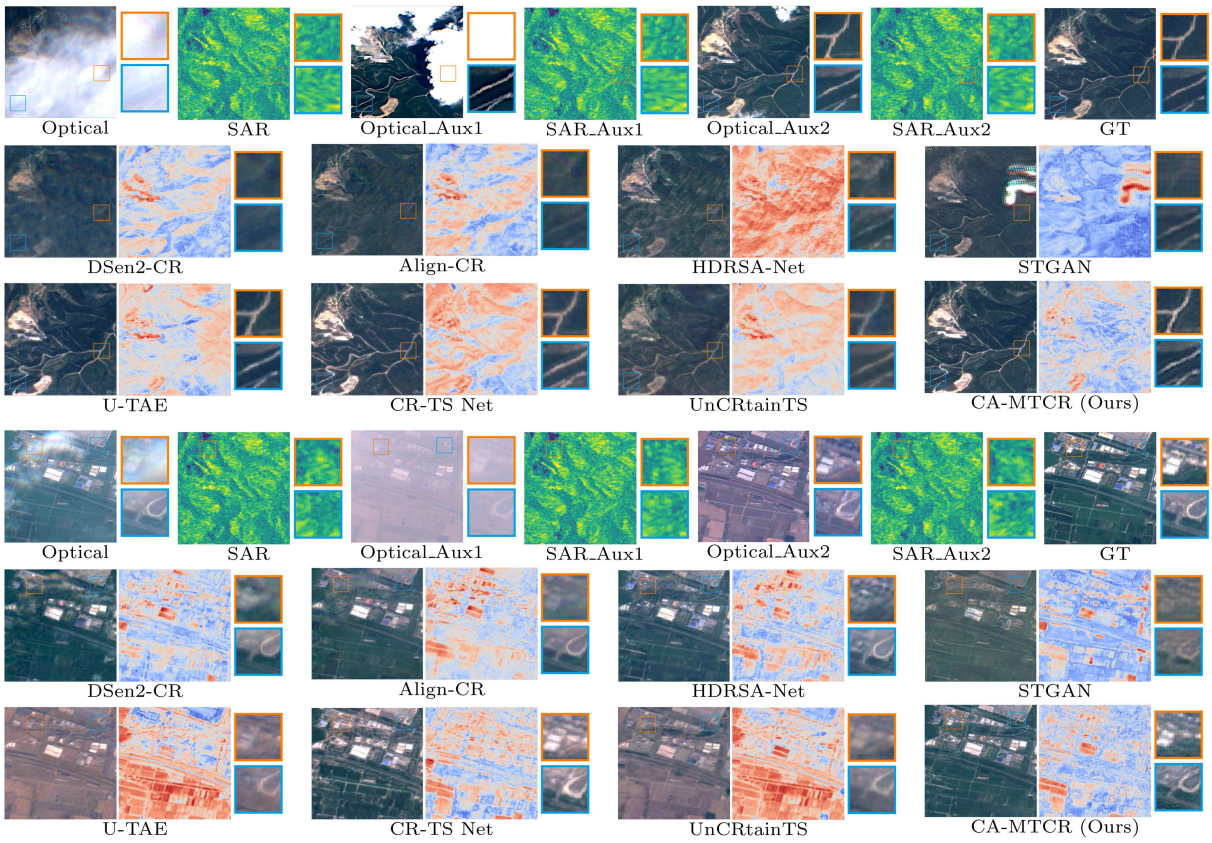

Figure 5 (Color online) Visualization of cloud removal results for two scenes in the SEN12MS-CR-TS dataset.

Table 2 Ablation study on the region-selective optical encoder and the change-aware multi-temporal fusion module on the SEN12MS-CR-TS dataset. ↑ indicates that higher values correspond to better performance, ↓ indicates that lower values correspond to better performance, and the best-performing results are highlighted in bold.

Method	RegS	CA	PSNR (dB) ↑	SAM (°) ↓	SSIM↑	MAE (ρ_{TOA}) ↓
Baseline			29.97	6.809	0.885	0.025
w/o RegS		✓	31.62 (+1.65)	5.184 (−1.625)	0.902 (+0.017)	0.020 (−0.005)
w/o CA	✓		30.16 (+0.19)	6.668 (−0.141)	0.885 (+0.000)	0.024 (−0.001)
CA-MTCR	✓	✓	**31.79** (**+1.82**)	**5.181** (**−1.628**)	**0.903** (**+0.018**)	**0.019** (**−0.006**)

reconstruction and degrades the visual homogeneity of the results. Our method pretrains the SAR encoder, enabling it to effectively suppress the adverse influence of speckle noise, thereby generating more uniform and visually coherent reconstructions in homogeneous areas.

4.3 Ablation study

The proposed CA-MTCR network improves the performance of multi-temporal cloud removal by optimizing the extraction of valuable information from mono-temporal observations and by strengthening multi-temporal fusion through the incorporation of change information detected from SAR images. In this section, we conduct ablation studies on the SEN12MS-CR-TS dataset to validate the effectiveness of each component, with the results shown in Table 2.

Mono-temporal feature extraction with region-selective encoding. CA-MTCR selectively aggregates non-local information from partially unobscured regions, mitigating the interference caused by redundant features in cloud-covered areas to enhance the feature representation of individual time steps. We validate its effectiveness by removing the region-selective encoding mechanism, i.e., retaining only the Transformer blocks in the region-selective optical encoder, denoted as "w/o RegS". It can be observed that the proposed CA-MTCR method, which excludes cloud-affected regions from feature encoding, achieves a performance gain of 0.17 dB in terms of PSNR, as it can more accurately capture ground object information, thereby benefiting the subsequent fusion process.

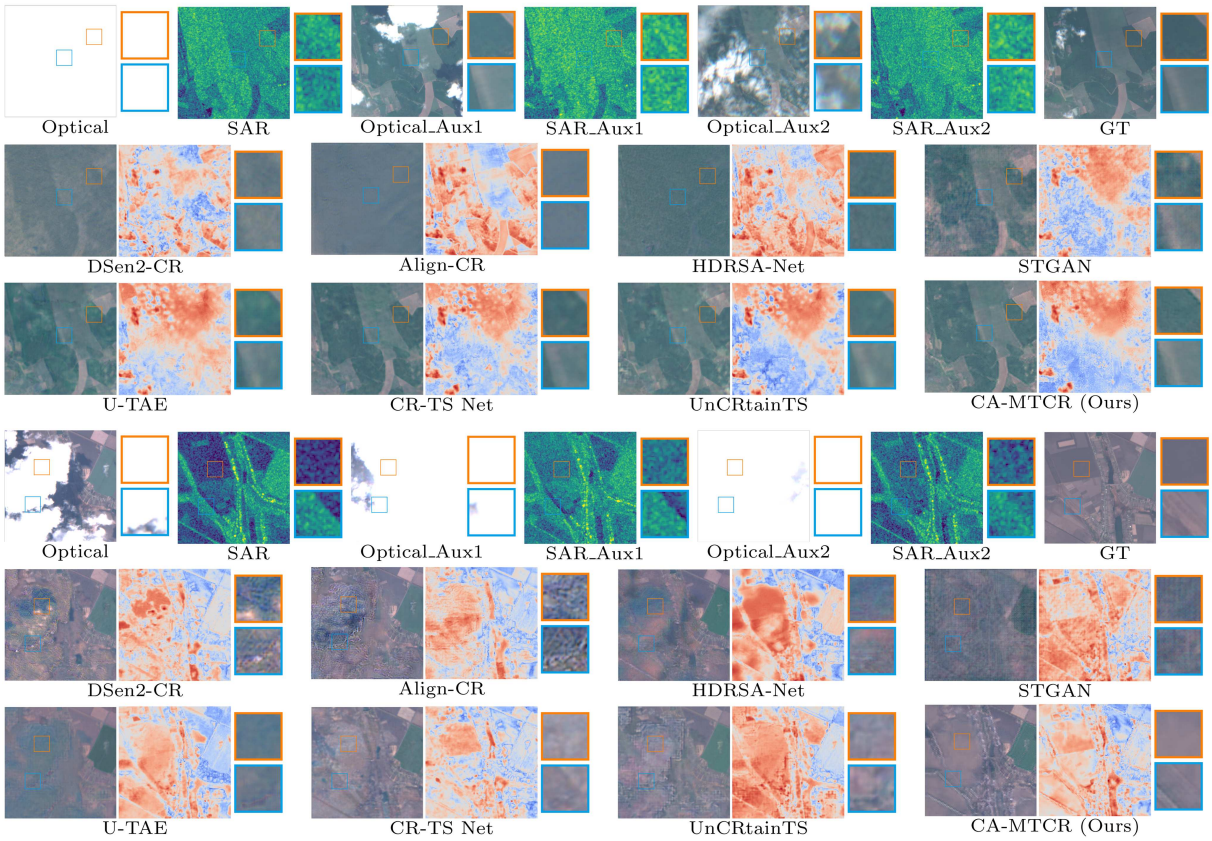

Figure 6 (Color online) Visualization of cloud removal results for two scenes in the AllClear dataset.

Multi-temporal fusion with change awareness. CA-MTCR optimizes the utilization of multi-temporal data by leveraging changes in ground objects detected from SAR images. To validate its superiority, we exclude change information when computing weights across time steps, denoted as "w/o CA". It can be seen that removing change awareness results in a noticeable performance decline, with a drop of 1.63 dB in PSNR, as the model struggles to effectively capture temporal variations, thereby reducing the accuracy of multi-temporal fusion.

4.4 Impact of temporal offset in auxiliary data

We further compare the proposed CA-MTCR networks with the best baseline, UnCRtainTS, using multi-temporal data with different offsets as auxiliary input. We evaluate the performance of cloud removal by integrating auxiliary data from time windows at different temporal distances from the current time step, and show the comparison results in terms of PSNR in Figure 7(a). We observe that the performance of UnCRtainTS progressively declines as the temporal distance between the auxiliary data and the target time step to be reconstructed increases. Similarly, for w/o CA, which does not utilize change information to optimize the use of multi-temporal data, the performance trends relative to temporal distance align with those of UnCRtainTS. As the temporal distance between the auxiliary data and the target time step grows, the mismatch between the data also increases, leading to reduced effectiveness in cloud removal. In contrast, CA-MTCR leverages SAR images to detect changes and guide the fusion of multi-temporal data, significantly mitigating performance degradation with larger temporal offsets. Notably, the cloud removal performance using auxiliary data from the time window 1–2 time steps distant from the target time step is comparable to that from the time window 3–4 time steps, as changes over shorter time intervals are typically minimal. However, due to the higher likelihood of cloud coverage in temporally closer data, the performance of the time window 1–2 time steps distant from the target time step may slightly underperform compared to time windows 3–4 when used as auxiliary input.

4.5 Impact of varying temporal data lengths

Theoretically, the more auxiliary data available, the more information can be utilized to compensate for cloud-covered regions at the target time step, thereby boosting reconstruction performance. We evaluate the proposed

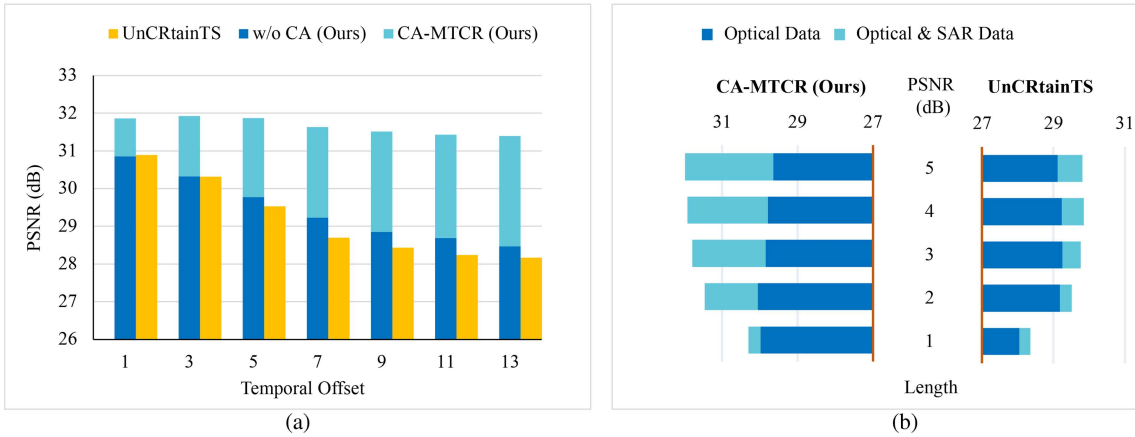

Figure 7 (Color online) (a) The results of UnCRtainTS and our proposed CA-MTCR using multi-temporal data with different offsets as auxiliary input on the SEN12MS-CR-TS dataset; (b) the results of UnCRtainTS and our proposed CA-MTCR using multi-temporal data—comprising either multi-temporal optical images or multi-temporal optical-SAR image pairs—of varying lengths as auxiliary input on the SEN12MS-CR-TS dataset.

CA-MTCR network against UnCRtainTS, using multi-temporal data—comprising either multi-temporal optical images or multi-temporal optical-SAR image pairs—of varying lengths as auxiliary input. The results are shown in Figure 7(b). CA-MTCR consistently outperforms UnCRtainTS across different auxiliary data lengths. For both methods, integrating SAR data significantly enhances their performance. Further, we find that when the length of the auxiliary data is zero, i.e., mono-temporal cloud removal, the benefits derived from SAR images are similar for both methods. However, when the length of the auxiliary data is non-zero, i.e., multi-temporal cloud removal, our proposed CA-MTCR achieves notably greater gains from SAR data. Specifically, as the length of the auxiliary data increases, the contribution from SAR data within CA-MTCR becomes progressively more crucial, primarily due to the heightened need to address disturbances caused by temporal inconsistencies. Our method strategically leverages SAR data to mitigate these inconsistencies, thus enhancing the utility of SAR data. Moreover, we observe that the performance of UnCRtainTS does not continuously improve with the increasing length of auxiliary time steps, because UnCRtainTS treats multi-temporal images uniformly, where more time steps introduce greater interference, resulting in decreased performance. Our method, when not utilizing SAR data, exhibits a performance trend similar to that of UnCRtainTS as the length of auxiliary data changes. However, when SAR data are available, our proposed CA-MTCR method can optimize the fusion of multi-temporal data, enabling more effective utilization of valuable information in the auxiliary data, thereby leading to consistent performance improvement.

5 Discussion

5.1 Performance under challenging conditions

In this work, our primary objective is to reconstruct cloud-free optical images at the target time step by leveraging multi-temporal observations. However, there exist particularly challenging scenarios where historical optical data provide little or no reference value.

Extreme cloud coverage. When the target as well as historical optical images are entirely obscured by clouds, as illustrated in the first and second scenes of Figure 8, optical observations fail to provide usable surface information. Our method reconstructs cloud-free imagery by exploiting fused representations of optical and SAR data, thereby relying on information embedded in SAR images to produce reasonable approximations of the underlying scenes. While SAR images provide valuable structural cues for reconstructing cloud-obscured regions, they inherently lack spectral attributes such as reflectance and color, which leads to inevitable degradation in the spectral fidelity of the reconstructed results.

Rapid land cover changes. When land cover undergoes rapid changes, historical observations can no longer serve as reliable references for reconstruction, as illustrated in the third and fourth scenes of Figure 8. In such cases, the best-performing baseline, UnCRtainTS, which does not explicitly account for inconsistencies in data collected at different times, tends to exploit clear regions from historical data, inadvertently introducing spurious details into the reconstructed results. By contrast, our method leverages SAR imagery to detect changes and guide the use of multi-temporal data, thereby effectively mitigating the adverse effects of temporal inconsistencies. By jointly

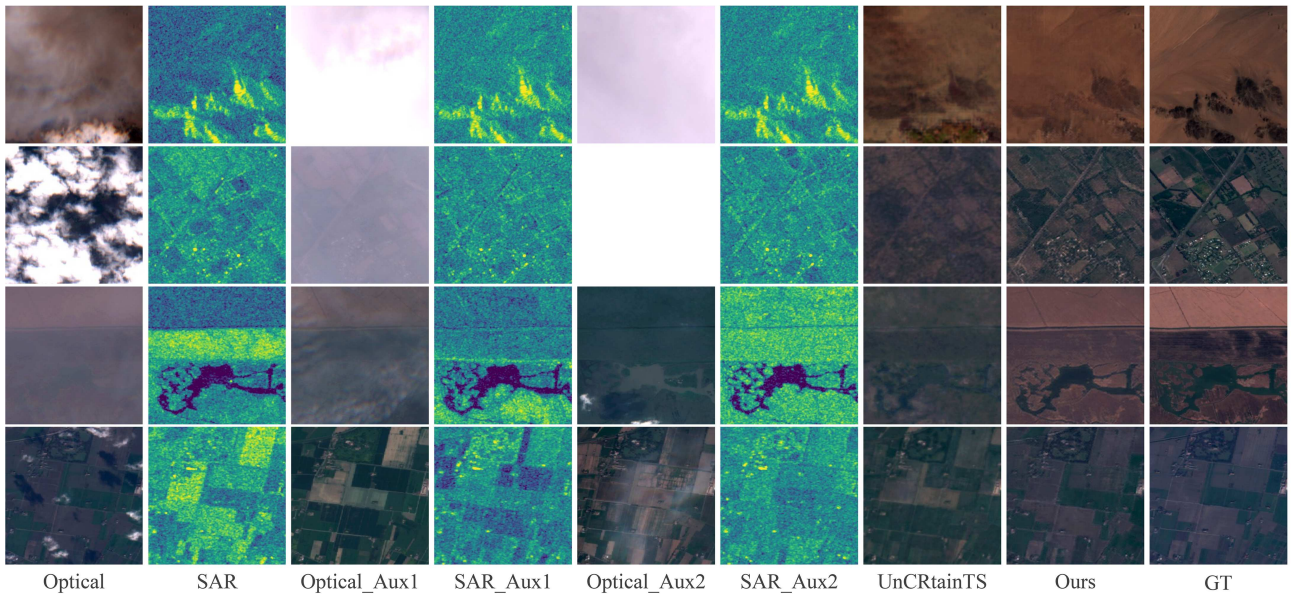

Figure 8 (Color online) Performance of UnCRtainTS and CA-MTCR (ours) under challenging conditions.

Table 3 Impact of resolution disparity between SAR and optical images on the SEN12MS-CR-TS dataset.

Method	PSNR (dB) ↑	SAM (°) ↓	SSIM ↑	MAE (ρ_{TOA}) ↓
w/o SAR	29.84	6.923	0.886	0.025
Original SAR	31.79	5.184	0.903	0.019
SAR downsampled to 1/5	31.68	5.225	0.902	0.020
SAR downsampled to 1/10	31.63	5.295	0.901	0.020

utilizing cloud-free portions of the optical image at the target time step together with structural information from the corresponding SAR observation, our method achieves reliable restoration of cloud-covered regions. When the optical image at the target time step is completely obscured by clouds, as illustrated in the third scene of Figure 8, the results resemble those observed under extreme cloud cover. In this case, no valuable spectral information is available, and the reconstructed image inevitably exhibits degradation in spectral fidelity.

5.2 Impact of resolution disparity between SAR and optical images

Multi-modal observations inherently exhibit disparities in spatial resolution, which may pose challenges to the stability of information integration. In the proposed framework, SAR data are primarily exploited to capture temporal change information for optimizing the utilization of multi-temporal observations, while the reconstruction of fine structural details is predominantly attributed to multi-temporal optical observations. Consequently, the impact of resolution mismatch is relatively limited. To further examine the robustness of our approach under resolution disparity, we conduct additional experiments in which SAR images are downsampled to 1/5 and 1/10 of their original resolution, with the results reported in Table 3. It can be observed that performance inevitably declines due to the loss of spatial detail. However, the performance degradation is relatively modest. For example, when the resolution of the SAR images is reduced to 1/5 of that of the optical images, the PSNR drops by only 0.11 dB. Notably, despite the presence of resolution disparities, the proposed method continues to yield substantial performance gains compared with scenarios where SAR data are not incorporated.

5.3 Potential extensions

Our method can effectively reconstruct cloud-free imagery at the target time step, offering strong potential for timely scene analysis. For time-critical applications such as disaster response, precision agriculture, or rapid environmental monitoring, the ability to deliver cloud-free observations in near real time is crucial. On a single NVIDIA 3090 GPU, without using any acceleration libraries, our framework achieves a processing speed of 8.45 FPS. Considering that current remote sensing video satellites (e.g., Jilin-1) typically operate at frame rates between 1 and 10 FPS, the computational efficiency of our method is well aligned with practical operational requirements.

6 Conclusion

In this paper, we propose CA-MTCR, a novel change-aware multi-temporal cloud removal method that leverages SAR images to capture temporal changes, optimizing multi-temporal data utilization to mitigate performance degradation caused by potential discrepancies between the target and auxiliary time steps. The accurate representation of valuable information from optical images, with the interference of cloud cover, is considered through a region-selective optical encoder to enhance the spectral fidelity of the reconstructed images. Experimental results on the SEN12MS-CR-TS and Allclear datasets demonstrate that our approach effectively improves the performance of multi-temporal cloud removal and exhibits robustness to variations in the length of auxiliary time steps and their offset relative to the target time step.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 62401406, U22B2011), Postdoctoral Fellowship Program of China Postdoctoral Science Foundation (CPSF) (Grant No. GZB20240562), China Postdoctoral Science Foundation (Grant No. 2024M762485), and Postdoctor Project of Hubei Province (Grant No. 2024HBBHJD076). The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

References

1 Reddy G P O. Satellite remote sensing sensors: principles and applications. In: Proceedings of Geospatial Technologies in Land Resources Mapping, Monitoring and Management, 2018. 21–43
2 Li K, Wan G, Cheng G, et al. Object detection in optical remote sensing images: a survey and a new benchmark. ISPRS J Photogramm Remote Sens, 2020, 159: 296–307
3 Zhao Q, Yu L, Du Z, et al. An overview of the applications of Earth observation satellite data: impacts and future trends. Remote Sens, 2022, 14: 1863
4 Sun X Q, Weng X X, Pang C, et al. Mitigating representation bias for class-incremental semantic segmentation of remote sensing images. Sci China Inf Sci, 2025, 68: 182301
5 Liu T Z, Hu B Y, Gu Y F, et al. An enhanced classification method based on adaptive multi-scale fusion for long-tailed multispectral point clouds. Sci China Inf Sci, 2025, 68: 182302
6 Hua Y, Zhu J, Li Q. PCINet: a prototype- and concept-based interpretable network for mutli-scene recognition. Int Arch Photogramm Remote Sens Spatial Inf Sci, 2024, XLVIII-1-2024: 265–270
7 Gawlikowski J, Ebel P, Schmitt M, et al. Explaining the effects of clouds on remote sensing scene classification. IEEE J Sel Top Appl Earth Obs Remote Sens, 2022, 15: 9976–9986
8 Xu F, Shi Y, Yang W, et al. CloudSeg: a multi-modal learning framework for robust land cover mapping under cloudy conditions. ISPRS J Photogramm Remote Sens, 2024, 214: 21–32
9 Pan J, Xu J, Yu X, et al. HDRSA-Net: Hybrid dynamic residual self-attention network for SAR-assisted optical image cloud and shadow removal. ISPRS J Photogramm Remote Sens, 2024, 218: 258–275
10 Ebel P, Meraner A, Schmitt M, et al. Multisensor data fusion for cloud removal in global and all-season Sentinel-2 imagery. IEEE Trans Geosci Remote Sens, 2020, 59: 5866–5878
11 Xu F, Shi Y, Ebel P, et al. Multimodal and multiresolution data fusion for high-resolution cloud removal: a novel baseline and benchmark. IEEE Trans Geosci Remote Sens, 2024, 62: 1–15
12 Li Y, Wei F, Zhang Y, et al. HS2P: Hierarchical spectral and structure-preserving fusion network for multimodal remote sensing image cloud and shadow removal. Inf Fusion, 2023, 94: 215–228
13 Li C, Liu X, Li S. Transformer meets GAN: cloud-free multispectral image reconstruction via multisensor data fusion in satellite images. IEEE Trans Geosci Remote Sens, 2023, 61: 1–13
14 Li X, Zhao X, Wang F, et al. HF-T2CR: high-fidelity thin and thick cloud removal in optical satellite images through SAR fusion. IEEE Trans Geosci Remote Sens, 2024, 62: 1–13
15 Xu F, Shi Y, Ebel P, et al. GLF-CR: SAR-enhanced cloud removal with global-local fusion. ISPRS J Photogramm Remote Sens, 2022, 192: 268–278
16 Wang Y, Zhang B, Zhang W, et al. Cloud removal with SAR-optical data fusion using a unified spatial-spectral residual network. IEEE Trans Geosci Remote Sens, 2024, 62: 1–20
17 Samadzadegan F, Toosi A, Dadrass Javan F. A critical review on multi-sensor and multi-platform remote sensing data fusion approaches: current status and prospects. Int J Remote Sens, 2025, 46: 1327–1402
18 Duan C, Belgiu M, Stein A. Efficient cloud removal network for satellite images using SAR-optical image fusion. IEEE Geosci Remote Sens Lett, 2024, 21: 1–5
19 Zhang Q, Yuan Q, Li J, et al. Thick cloud and cloud shadow removal in multitemporal imagery using progressively spatio-temporal patch group deep learning. ISPRS J Photogramm Remote Sens, 2020, 162: 148–160
20 Stucker C, Garnot V S F, Schindler K. U-TILISE: a sequence-to-sequence model for cloud removal in optical satellite time series. IEEE Trans Geosci Remote Sens, 2023, 61: 1–16
21 Zhou H, Kao C H, Phoo C P, et al. AllClear: a comprehensive dataset and benchmark for cloud removal in satellite imagery. 2024. ArXiv:2410.23891
22 Christopoulos D, Ntouskos V, Karantzalos K. CloudTran++: improved cloud removal from multi-temporal satellite images using axial transformer networks. Remote Sens, 2025, 17: 86
23 Sarukkai V, Jain A, Uzkent B, et al. Cloud removal from satellite images using spatiotemporal generator networks. In: Proceedings of IEEE/CVF Winter Conference on Applications of Computer Vision, 2020. 1796–1805
24 Ebel P, Xu Y, Schmitt M, et al. SEN12MS-CR-TS: a remote-sensing data set for multimodal multitemporal cloud removal. IEEE Trans Geosci Remote Sens, 2022, 60: 1–14
25 Ebel P, Garnot V S F, Schmitt M, et al. UnCRtainTS: uncertainty quantification for cloud removal in optical satellite time series. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2023. 2086–2096
26 Li W, Yang W, Liu T, et al. Predicting gradient is better: exploring self-supervised learning for SAR ATR with a joint-embedding predictive architecture. ISPRS J Photogramm Remote Sens, 2024, 218: 326–338
27 Li W, Yang W, Hou Y, et al. SARATR-X: toward building a foundation model for SAR target recognition. IEEE Trans Image Process, 2025, 34: 869–884
28 Chen Z Y, Li Y H, Hu C, et al. Repeat-pass space-surface bistatic SAR tomography: accurate imaging and first experiment. Sci China Inf Sci, 2024, 67: 192304
29 Lehtinen J, Munkberg J, Hasselgren J, et al. Noise2noise: learning image restoration without clean data. 2018. ArXiv:1803.04189
30 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

31 Cheng Q, Shen H, Zhang L, et al. Missing information reconstruction for single remote sensing images using structure-preserving global optimization. IEEE Signal Process Lett, 2017, 24: 1163–1167
32 Liang S, Fang H, Chen M. Atmospheric correction of Landsat ETM+ land surface imagery. I. Methods. IEEE Trans Geosci Remote Sens, 2001, 39: 2490–2498
33 Shen H, Li X, Zhang L, et al. Compressed sensing-based inpainting of Aqua Moderate Resolution Imaging Spectroradiometer band 6 using adaptive spectrum-weighted sparse Bayesian dictionary learning. IEEE Trans Geosci Remote Sens, 2013, 52: 894–906
34 Xu M, Pickering M, Plaza A J, et al. Thin cloud removal based on signal transmission principles and spectral mixture analysis. IEEE Trans Geosci Remote Sens, 2015, 54: 1659–1669
35 Enomoto K, Sakurada K, Wang W, et al. Filmy cloud removal on satellite imagery with multispectral conditional generative adversarial nets. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2017. 48–56
36 Li J, Wu Z, Hu Z, et al. 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
37 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 IEEE International Geoscience and Remote Sensing Symposium, 2018. 1726–1729
38 Gao J, Yuan Q, Li J, et al. Cloud removal with fusion of high resolution optical and SAR images using generative adversarial networks. Remote Sens, 2020, 12: 191
39 Meraner A, Ebel P, Zhu X X, et al. 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
40 Lin C H, Tsai P H, Lai K H, et al. Cloud removal from multitemporal satellite images using information cloning. IEEE Trans Geosci Remote Sens, 2012, 51: 232–241
41 Dai P, Ji S, Zhang Y. Gated convolutional networks for cloud removal from Bi-temporal remote sensing images. Remote Sens, 2020, 12: 3427
42 Zheng W J, Zhao X L, Zheng Y B, et al. Spatial-spectral-temporal connective tensor network decomposition for thick cloud removal. ISPRS J Photogramm Remote Sens, 2023, 199: 182–194
43 Chen Z, Zhang P, Zhang Y, et al. Thick cloud removal in multi-temporal remote sensing images via frequency spectrum-modulated tensor completion. Remote Sens, 2023, 15: 1230
44 Zhang K, Nie H, Li W, et al. A multitemporal remote sensing thick cloud removal network based on implicit reconstruction. Int J Remote Sens, 2025, 46: 1574–1593
45 Gao G, Gu Y. Multitemporal Landsat missing data recovery based on tempo-spectral angle model. IEEE Trans Geosci Remote Sens, 2017, 55: 3656–3668
46 Czerkawski M, Upadhyay P, Davison C, et al. Deep internal learning for inpainting of cloud-affected regions in satellite imagery. Remote Sens, 2022, 14: 1342
47 Sandler M, Howard A, Zhu M, et al. MobileNetV2: inverted residuals and linear bottlenecks. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018. 4510–4520
48 Dineshkumar R, Pushpavalli R, Obaid S, et al. Deep separable dilated convolutional neural network for cloud detection and removal in satellite imagery. In: Proceedings of International Conference on Software, Systems and Information Technology, 2024. 1–5
49 Anandakrishnan J, Sundaram V M, Paneer P. CERMF-Net: a SAR-optical feature fusion for cloud elimination from Sentinel-2 imagery using residual multiscale dilated network. IEEE J Sel Top Appl Earth Obs Remote Sens, 2024, 17: 11741–11749
50 Argenti F, Lapini A, Bianchi T, et al. A tutorial on speckle reduction in synthetic aperture radar images. IEEE Geosci Remote Sens Mag, 2013, 1: 6–35
51 Van den Oord A, Vinyals O, Kavukcuoglu K, et al. Neural discrete representation learning. 2017. ArXiv:1711.00937
52 Garnot V S F, Landrieu L, Giordano S, et al. Satellite image time series classification with pixel-set encoders and temporal self-attention. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020. 12325–12334
53 Vincent E, Saroufim M, Chemla J, et al. Detecting looted archaeological sites from satellite image time series. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2025. 2296–2307
54 Thoreau R, Marsocci V, Derksen D. Parameter-efficient adaptation of geospatial foundation models through embedding deflection. 2025. ArXiv:2503.09493
55 Garnot V S F, Landrieu L. Panoptic segmentation of satellite image time series with convolutional temporal attention networks. 2021. ArXiv:2107.07933