Exploring the Potential of SAR Data for Cloud Removal in Optical Satellite Imagery

Fang Xu\textsuperscript{a,b}, Yilei Shi\textsuperscript{c}, Patrick Ebel\textsuperscript{b}, Lei Yu\textsuperscript{a}, Gui-Song Xia\textsuperscript{d}, Wen Yang\textsuperscript{a,e} and Xiao Xiang Zhu\textsuperscript{a,e,*}

\textsuperscript{a}School of Electronic Information, Wuhan University, Wuhan, 430072, China
\textsuperscript{b}Data Science in Earth Observation, Technical University of Munich, Munich, 80333, Germany
\textsuperscript{c}Remote Sensing Technology, Technical University of Munich, Munich, 80333, Germany
\textsuperscript{d}School of Computer Science, Wuhan University, Wuhan, 430072, China
\textsuperscript{e}Remote Sensing Technology Institute, German Aerospace Center, Wessling, 82234, Germany

\textbf{ARTICLE INFO}

\begin{itemize}
  \item Keywords: Cloud removal
  \item Data fusion
  \item SAR
  \item Transformer
\end{itemize}

\textbf{ABSTRACT}

The challenge of the cloud removal task can be alleviated with the aid of Synthetic Aperture Radar (SAR) images that can penetrate cloud cover. However, the large domain gap between optical and SAR images as well as the severe speckle noise of SAR images may cause significant interference in SAR-based cloud removal, resulting in performance degeneration. In this paper, we propose a novel global-local fusion based cloud removal (GLF-CR) algorithm to leverage the complementary information embedded in SAR images. Exploiting the power of SAR information to promote cloud removal entails two aspects. The first, global fusion, guides the relationship among all local optical windows to maintain the structure of the recovered region consistent with the remaining cloud-free regions. The second, local fusion, transfers complementary information embedded in the SAR image that corresponds to cloudy areas to generate reliable texture details of the missing regions, and uses dynamic filtering to alleviate the performance degradation caused by speckle noise. Extensive evaluation demonstrates that the proposed algorithm can yield high quality cloud-free images and performs favorably against state-of-the-art cloud removal algorithms.

\section{1. Introduction}

Earth observation through satellites plays a vital role in understanding the world, and has attracted attention from a wide range of communities (Xia et al., 2018; Requena-Mesa et al., 2021; Girard et al., 2021). However, optical satellite images are often contaminated by clouds, which obstruct the view of the surface underneath, as shown in Fig. 1a. A study conducted by the MODIS instrument shows that clouds cover around 67\% of the earth’s surface (King et al., 2013). Thus, cloud removal becomes an indispensable preprocessing step for applications relying on data streams of continuous monitoring (Ebel et al., 2021). Due to the absence of textures in cloud-covered regions, the task of cloud removal is severely ill-posed. Benefiting from Synthetic Aperture Radar (SAR) (Bamler, 2000) (as shown in Fig. 1b), which is not affected by clouds due to its advantage of strong penetrability and inherently reflects the geometrical shapes of ground objects, the challenge of cloud removal can be essentially alleviated. However, the recovery of high-quality cloud-free images with the aid of SAR images is nevertheless a challenging problem due to the following issues:

- **Domain Gap.** SAR and optical images reveal different characteristics of observed objects due to their different imaging mechanisms, and thus a large domain gap exists between them (Schmitt et al., 2017; Liu and Lei, 2018). Transferring the complementary information from a SAR image to compensate for the missing information in cloudy regions is non-trivial.

- **Speckle Noise.** SAR images exhibit bright and dark pixels, i.e., speckle noise, which is uneven, even for homogeneous regions (Zhu et al., 2021; Yu et al., 2018). Moreover, the speckle noise usually exists in the same wave front as the surface information of the target. This undesirable effect leads to performance degradation on reconstruction (Fuentes Reyes et al., 2019).

A few SAR-based cloud removal methods to learn to transfer the concatenation of multi-modal images to cloud-free images have been proposed (Gao et al., 2020; Meranner et al., 2020; Ebel et al., 2021). However, the pixel-to-pixel translation does not take into account the long-range varying contextual information of the cloud-free regions, leading to texture and structure discrepancies. Furthermore, this concatenation method only partially explores the interactions or correlations between optical and SAR data, in which complementary information cannot be effectively transferred. Moreover, simply stacking multi-modal images is susceptible to speckle noise, which hinders the cloud removal performance.

As a consequence, there is an urgent need to develop new cloud removal methods to utilize sufficient information from SAR, in order to improve cloud removal performance. There is no doubt that SAR images help to recover texture details by compensating for the missing information in cloudy regions. In addition, since a SAR image is not obscured by clouds,
it contains reliable global contextual information that can provide valuable guidance for capturing global interactions between contexts to maintain global consistency with the remaining cloud-free regions.

In this paper, we propose a novel global-local fusion-based cloud removal (GLF-CR) algorithm that incorporates the contributions of SAR images described above. Specifically, it contains two parallel backbones developed for optical and SAR image representation learning, where SAR features are used in a hierarchical manner to compensate for the loss of information. Inspired by Transformer architectures (Vaswani et al., 2017) that can capture global interactions between contexts, we propose a SAR-guided global context interaction (SGCI) block in which SAR features are used to guide the interactions of global optical feature. Furthermore, a SAR-based local feature compensation (SLFC) block is proposed to transfer complementary information from the corresponding regions in the SAR features to the optical features, where dynamic filtering is used to handle speckle noise. Consequently, the proposed algorithm can generate knowledgeable features with comprehensive information, thereby yielding high-quality cloud-free images.

In a nutshell, the contributions of our work are threefold:

- We propose a novel SAR-based cloud removal algorithm, GLF-CR. It incorporates SAR’s contribution to restoring reliable texture details and maintaining global consistency, thus enabling the region occluded by cloud cover to be effectively reconstructed.

- We propose a SAR-guided global context interaction (SGCI) block, in which the SAR feature is used to guide the global interactions between contexts in order to ensure that the structure of the recovered cloud-free region is consistent with the remaining cloud-free regions.

- We propose a SAR-guided local feature compensation (SLFC) block to enhance the transference of complementary information embedded in the SAR image while avoiding the influence of speckle noise, and thus generate more reliable texture details.

2. Related Work

Cloud Removal. Cloud removal aims to reconstruct the missing information caused by clouds in optical satellite imagery. Early attempts address this problem by assuming the corrupted regions and the remaining regions share the same statistical and geometrical structures. They view cloud removal as an inpainting task and use the information around the corrupted regions to predict the missing data (Chan and Shen, 2001; Maalouf et al., 2009; Singh and Komodakis, 2018). These methods fail to make accurate inferences when the corrupted region occupies a large portion of the image. To mitigate this problem, a series of studies make use of multispectral data to recover the missing information (Xu et al., 2015; Shen et al., 2013; Enomoto et al., 2017). For example, McGANs (Enomoto et al., 2017) utilizes additional near-infrared (NIR) images, which process higher penetrability through clouds, to improve visibility. However, as the thickness of clouds increases, all the land signals in the optical bands are obstructed. Consequently, multitemporal-based approaches have been proposed to restore the missing information with data from other time periods (Zhang et al., 2021; Shen et al., 2019; Ebel et al., 2022). However, when encountering continual cloudy days, cloud-free reference data from an adjacent period is largely unavailable.

Synthetic Aperture Radar (SAR) images are cloud-penetrable and thus provide missing information due to optically impenetrable clouds (Bamler, 2000). There is promising potential in SAR-to-optical image translation. Some researchers have tried to generate optical images directly from SAR (Fuentes Reyes et al., 2019; Bermudez et al., 2018). However, since SAR lacks spectrally resolved measurements, there are domain-specific potentials and
peculiarities that cannot be compensated. It is challenging to guarantee the quality of the generated optical image translated from a SAR image. Recently, a few works have explored the means of SAR-optical data fusion, exploiting the synergistic properties of the two imaging systems to guide cloud removal. Meraner et al. concatenate the SAR image to the input optical image and use a deep residual neural network to predict the target cloud-free optical image (Meraner et al., 2020). Gao et al. utilize a two-step approach, first translating the SAR image into a simulated optical image, and then concatenating the simulated optical image, the SAR image, and the optical image corrupted by clouds to reconstruct the corrupted regions using the generative adversarial network (GAN) (Gao et al., 2020). Experiments have verified the usefulness of SAR-optical data fusion, but its gain is limited because the concatenation approach has limited ability to utilize the complementary information from the SAR image. To boost the gain that comes with the additional SAR information, we propose a novel cloud removal algorithm, GLF-CR, which incorporates SAR’s contribution to restoring reliable texture details and maintaining global consistency to compensate for information loss in cloudy regions.

**Image Restoration.** Cloud removal is essentially an image restoration task in which a high-quality clean image is reconstructed from a low-quality, degraded counterpart. Recent advances in image restoration follow convolutional neural network (CNN), and numerous CNN-based models have been proposed to improve restoration performance (Zhang et al., 2018a, 2020; Wang et al., 2021). Global context plays an important role in local pixel-wise recovery. However, convolution is not effective for long-range dependency modeling under the principle of local processing (Liang et al., 2021). To ensure visually consistent restoration results, a series of works focus on the attention mechanism to obtain global dependency information. Wang et al. exploit a two-round four-directional IRNN architecture to accumulate global contextual information (Wang et al., 2019). Zheng et al. introduce a short+long term attention layer to ensure appearance consistency in the image domain. Recently, Transformer that employs a self-attention mechanism to capture global interactions between contexts (Liu et al., 2021b) has been proposed and shows promising performance in image restoration (Liang et al., 2021). However, these methods mostly focus on reconstructing a high-quality clean image from a single low-quality degraded counterpart without auxiliary data. In contrast, our proposed model accomplishes high-quality image restoration by integrating complementary information.

Most existing cloud removal methods are carried out by extending the input channels of the popular CNN architectures. For example, McGAN (Enomoto et al., 2017) extends the input channels of the conditional Generative Adversarial Networks (cGANs) so that they are compatible with multispectral images. DSen2CR (Meraner et al., 2020) is derived from the EDSR network (Lim et al., 2017), and concatenates the SAR’s channels and the other channels of the input optical image as input. These architectures are usually designed for tasks like super-resolution and motion deblurring, where the local information from the original low-quality image is only partially lost. For the cloud removal task, all the local information in the area covered by thick clouds is missing because the clouds completely corrupt the reflectance signal. Thus, the cloud removal methods extended from these architectures have limited ability to fully utilize the spatial consistency between the cloudy and the neighboring cloud-free regions. In comparison, the architecture presented in this work is designed to integrate the global context information under the guidance of the SAR image.

**Multi-Modal Data Fusion.** Commonly used fusion strategies include element-wise multiplication/addition or concatenation between different types of features (Xu et al., 2021; Sun et al., 2019; Fu et al., 2020); this multi-modal data fusion yields limited performance gain (Wu and Han, 2018; Audebert et al., 2018; Liu et al., 2021a). To better exploit the complementary information of the auxiliary data, Hazirbas et al. propose FuseNet for semantic segmentation. FuseNet contains two branches to extract features from the RGB and depth images, and constantly fuses them via element-wise summation (Hazirbas et al., 2016). Liu et al. propose an information aggregation distribution module for crowd counting, which consists of two branches for modality-specific representation learning (i.e., RGB and thermal image) and an additional branch for modality-shared representation learning. It dynamically enhances the modality-shared and modality-specific representations with a dual information propagation mechanism (Liu et al., 2021a). These methods increase the utilization of complementary information of auxiliary data. Nevertheless, little consideration has been given to SAR-optical data fusion for cloud removal, the specific challenges of which are addressed and resolved in this work.

### 3. Problem Statement

Given a cloudy image \( I \) defined over \( X \cong C + \Theta \) with \( C \) and \( \Theta \) respectively denoting the **cloud-covered** and **cloud-free** regions, the task of cloud removal aims at restoring the cloud-covered region of the image, i.e., \( I_C \). Generally, this task is severely ill-posed due to the missing information caused by clouds in optical satellite observations.

**Inpainting.** The basic strategy is to infer the cloud-covered region \( I_C \) from the cloud-free part \( I_\Theta \), and thus it can be considered as inpainting task, i.e.,

\[
I_C = \text{F}_{\text{INP}}(I_\Theta; S(I)),
\]

where \( \text{F}_{\text{INP}} \) is an inpainting operator conditioned by latent structures of the whole image, i.e., \( S(I) \). Specifically, \( S(I) \) represents priors of images, e.g., smoothness, non-local similarities, or learned features embedding from data, with which the task of cloud removal is tractable. However, latent structures of \( S(I) \) are not generally holistic or are even unavailable in a cloud removal task when the cloud-covered
Fig. 2: Overview of the proposed global-local fusion based cloud removal (GLF-CR) algorithm. It is a two-stream network in which the SAR feature is hierarchically fused into the optical feature to compensate for information loss in cloudy areas. Exploiting the power of SAR information to promote cloud removal entails two aspects: global fusion, to guide the relationship among all local optical windows based on the SAR-guided global context interaction (SGCI) block; and local fusion, to transfer the SAR feature corresponding to cloudy areas based on the SAR-based local feature compensation (SLFC) block.

The main obstacles to boosting cloud removal performance are two-fold.

- The network should effectively transfer the complementary information from SAR image $B_c$ to the optical image while overcoming the influence of its speckle noise to generate reliable texture details.
- The surface information from the cloud-free region $I_O$ should be considered to maintain the structure of the recovered cloud-free region consistent with the remaining cloud-free regions.

Global-Local Fusion. Thus our task is to develop an operator $F_{fusion}$ conditioned by both the inherent relationship between SAR and optical images and latent structures of the whole image, i.e.,

$$I_C = F_{fusion}(I_O, B_c; S(I/B), R(B, I)).$$  \hspace{1cm} (3)

where $S(I/B)$ is the non-local context information of the cloudy image learned under the guidance of SAR image. Since the SAR image is not affected by cloud cover, it can provide valuable guidance for capturing global interactions between contexts, so as to maintain the structure of the recovered cloud-free region consistent with the remaining cloud-free regions. $R(B, I)$ in $F_{fusion}$ is different from its counterpart in $F_{TRF}$, which incorporates the information of the SAR image by stacking its channels to the optical image. We propose instead a more effective fusion strategy to transfer the complementary information from the corresponding region in the SAR image, so as to generate more reliable texture details.

4. Method

4.1. Overview

The overall framework of the proposed GLF-CR algorithm is illustrated in Fig. 2. It is a two-stream network in which the SAR feature is hierarchically fused into the optical feature to compensate for information loss in cloudy regions. Exploiting the power of SAR information to promote cloud removal entails two aspects: global fusion, to guide the relationship among all local optical windows with the SGCI block; and local fusion, to transfer the SAR feature corresponding to cloudy areas with the SLFC block.
Specifically, given a cloudy image $I$ and its corresponding SAR image $B$ as input, we first feed them into different branches to extract modality-specific features $\hat{F}^0_{\text{opt}}$ and $\hat{F}^0_{\text{sar}}$ with the shallow feature extraction (SFE) block,

$$\hat{F}^0_{\text{opt}} = H_{\text{SFE}_{\text{opt}}}(I), \hat{F}^0_{\text{sar}} = H_{\text{SFE}_{\text{sar}}}(B),$$  \hspace{1cm} (4)

where $H_{\text{SFE}_{\text{opt}}}(\cdot)$ and $H_{\text{SFE}_{\text{sar}}}(\cdot)$ denote the functions to extract the shallow features of the cloudy image and the SAR image, respectively. Then, we feed $\hat{F}^0_{\text{opt}}$ and $\hat{F}^0_{\text{sar}}$ into $D$ functions composed from the SGCI and SLFC block to obtain knowledgable features with comprehensive information. More specifically, we obtain the intermediate features $\{\hat{F}^1_{\text{opt}}, \hat{F}^1_{\text{sar}}\}, \{\hat{F}^2_{\text{opt}}, \hat{F}^2_{\text{sar}}\},..., \{\hat{F}^D_{\text{opt}}, \hat{F}^D_{\text{sar}}\}$ as

$$\hat{F}^i_{\text{opt}}, \hat{F}^i_{\text{sar}} = H_{\text{SLFC}}(H_{\text{SGCI}}(\hat{F}^{i-1}_{\text{opt}}, \hat{F}^{i-1}_{\text{sar}})),$$  \hspace{1cm} (5)

where $H_{\text{SGCI}}(\cdot)$ and $H_{\text{SLFC}}(\cdot)$ denote the functions of the SGCI block and the SLFC block, respectively. The purpose of the SGCI block is local feature extraction and cross-window feature interaction, where the SAR feature is used to guide the relationship among all local optical windows. Each SGCI block is followed by an SLFC block, which is designed to fuse the complementary information from the corresponding area in a SAR image into the optical feature of a cloudy area. More details about these two blocks will be given in Secs. 4.2 and 4.3. Finally, we reconstruct the high-quality cloud-free image $I_C$ by aggregating all the intermediate optical features,

$$I_C = I + H_{\text{IR}}([\hat{F}^1_{\text{opt}}, \hat{F}^2_{\text{opt}}, ..., \hat{F}^D_{\text{opt}}]).$$  \hspace{1cm} (6)

where $H_{\text{IR}}$ denotes the function of cloud-free image reconstruction, and $[\hat{F}^1_{\text{opt}}, \hat{F}^2_{\text{opt}}, ..., \hat{F}^D_{\text{opt}}]$ refers to the concatenation of the intermediate optical features.

### 4.2. SAR-Guided Global Context Interaction

The SGCI block, whose detail is shown in Fig. 3, has two parallel streams for the input optical and SAR features. Each stream adopts dense connections in an approach similar to the residual dense block (RDB) (Zhang et al., 2018b), which is able to extract abundant local features via dense connected convolutional layers. As previously mentioned, SAR image clearly contributes to compensating for the missing information about cloudy regions, but not for the specific properties of optical images. Nevertheless, the cloud-free regions are conducive to the specific properties. The use of global texture information is necessary for the cloud removal task. Inspired by Transformer’s ability to efficiently propagate information across the entire image to accumulate long-range varying contextual information, we add a Swin Transformer layer (STL) (Liu et al., 2021b) after each local convolutional layer for cross-window feature interaction.

The STL first partitions the input feature into non-overlapping $M \times M$ windows, then computes the standard self-attention separately for each window. Specifically, a local window optical/SAR feature $X^{i}_{\text{opt}}/X^{i}_{\text{sar}} \in \mathbb{R}^{M^2 \times C}$ is linearly transformed to query $Q^{i}_{\text{opt}}/Q^{i}_{\text{sar}} \in \mathbb{R}^{M^2 \times d}$, key

Fig. 3: Detail of the SAR-guided global context interaction (SGCI) block.

$K^{i}_{\text{opt}}/K^{i}_{\text{sar}} \in \mathbb{R}^{M^2 \times d}$, and value $V^{i}_{\text{opt}}/V^{i}_{\text{sar}} \in \mathbb{R}^{M^2 \times d}$, where $d$ is the dimension of the query or key. The attention weight matrix is computed as follows:

$$M^{i}_{\text{opt}} = \frac{Q^{i}_{\text{opt}}K^{T}_{\text{opt}}}{\sqrt{d}} + B, M^{i}_{\text{sar}} = \frac{Q^{i}_{\text{sar}}K^{T}_{\text{sar}}}{\sqrt{d}} + B,$$  \hspace{1cm} (7)

where $B$ is the learnable relative positional encoding. The essence of this attention matrix is the weight of a particular region that is absorbing information from other regions. For a cloudy region, due to the information loss, it is difficult to estimate its interactions with cloud-free regions. For the same region in a SAR image, its interactions with other regions can be estimated easily, which provides valuable guidance for the interactions between cloud-free and cloudy regions in the optical image. Thus, we transfer the attention map of the SAR image to refine the attention map of the optical image, i.e., we use $M_{\text{sar}}$ to improve $M_{\text{opt}}$. We first obtain the attention map of the optical and SAR features $M_{\text{opt}}$ and $M_{\text{sar}}$ by Eq. (7). Then we compute the difference between $M_{\text{opt}}$ and $M_{\text{sar}}$ and obtain $M_{\text{res}}$. Finally, we apply a gating function to adaptively refine $M_{\text{opt}}$:

$$\tilde{M}_{\text{opt}} = M_{\text{opt}} + M_{\text{res}} \odot G(M_{\text{res}}),$$  \hspace{1cm} (8)

where $G(\cdot)$ is the gating function fed with the residual term $M_{\text{res}}$ and $\odot$ denotes the element-wise multiplication operation. The optical and SAR output are computed as:

$$Y^{i}_{\text{opt}} = \text{Softmax}(\tilde{M}^{i}_{\text{opt}})V^{i}_{\text{opt}}, Y^{i}_{\text{sar}} = \text{Softmax}(M^{i}_{\text{sar}})V^{i}_{\text{sar}}.$$  \hspace{1cm} (9)

This module considers the relationship among all local window optical features under the guidance of the SAR feature, denoted in this paper as global fusion.
4.3. SAR-based Local Feature Compensation

The detail of the SLFC block is shown in Fig. 4. Because the SAR image is corrupted by severe speckle noise, we utilize dynamic filtering for SAR features before information transference. Standard convolution filters are shared across all pixels in an image, while the dynamic filters vary from pixel to pixel. Therefore, the dynamic filters can handle the spatial variance issue (Jia et al., 2016; Zhou et al., 2019), thus helping to suppress the spatially unevenly speckle noise. Specifically, we dynamically generate a filter for each position in the SAR feature using the Dynamic Filter Generation (DFG) module. The DFG module takes the concatenation of the optical and SAR features $\text{Concat}(F^i_{\text{opt}}, F^i_{\text{sar}}) \in \mathbb{R}^{H \times W \times C}$ as input. The dimension of the generated filter $F^i \in \mathbb{R}^{H \times W \times C k^2}$ is reshaped into a five-dimensional filter. Then, for each position $(h, w, c)$ in the SAR feature $F^i_{\text{sar}} \in \mathbb{R}^{H \times W \times C}$, a specific local filter $F^i(h, w, c) \in \mathbb{R}^{k \times k \times k}$ is applied to the region centered around $F^i_{\text{sar}}(h, w, c)$ as

$$\hat{F}^i_{\text{sar}}(h, w, c) = F^i(h, w, c) * F^i_{\text{sar}}(h, w, c), \quad (10)$$

where $*$ denotes the convolution operation.

After transforming the extracted SAR feature $F^i_{\text{sar}}$ using the dynamic filter to improve tolerance of speckle noise, we propagate the complementary information from the SAR feature to refine the optical feature, in the same way that the attention map is refined. We compute the difference between the optical and SAR features to obtain the residual information $F^i_{x-o} = \hat{F}^i_{\text{sar}} - F^i_{\text{opt}}$, and apply a gating function to transfer the complementary information,

$$\tilde{F}^i_{\text{opt}} = F^i_{\text{opt}} + F^i_{x-o} \odot G(F^i_{x-o}). \quad (11)$$

To better exploit interactions among elements of the optical and SAR features for a further performance gain, we adopt a dual information propagation mechanism, i.e., updating the SAR feature as well. We compute the difference between the SAR feature and the updated optical feature $F^i_{o-s} = \tilde{F}^i_{\text{opt}} - \hat{F}^i_{\text{opt}}$, and also propagate the information through use of a gating function,

$$\tilde{F}^i_{\text{sar}} = F^i_{\text{sar}} + F^i_{o-s} \odot G(F^i_{o-s}). \quad (12)$$

The enhanced optical and SAR features are then introduced into the next SGCI for further representation learning. This module considers the information transference between local features, denoted as local fusion in this paper.

5. Experiments

5.1. Experimental Settings

Dataset and Metrics. Our experiments are conducted on the large-scale dataset SEN12MS-CR (Ebel et al., 2021), which is built from freely available data acquired by the Sentinel satellites in the Copernicus program. The dataset contains 122,218 samples from 169 non-overlapping regions of interest (ROI) distributed over all inhabited continents during all meteorological seasons. Each sample consists of a triplet of an orthorectified, geo-referenced Sentinel-1 dual-pol SAR image, a Sentinel-2 cloud-free multi-spectral image, and a cloud-covered Sentinel-2 multi-spectral image where the observations of cloud-free and cloud-covered images are close in time. The size of each image is $256 \times 256$ pixels. The VV and VH polarizations of the SAR images are clipped to values $[-25, 0]$ and $[-32.5, 0]$, and rescaled to the range $[0, 1]$. All bands of the optical images are clipped to values $[0, 10000]$, and rescaled to the range $[0, 1]$ as well. The results of cloud removal are evaluated with the peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM), spectral angle mapper (SAM), and mean absolute error (MAE).

Implementation Details. Our proposed GLF-CR network is implemented using publicly available Pytorch and trained in an end-to-end manner supervised by L1 loss on 4 NVIDIA TITAN V GPUs. We implement the gating functions in Secs. 4.2 and 4.3 by employing a convolution layer as well as a Softmax layer, and the dynamic filter generation (DFG) module in Sec. 4.3 is constituted by a convolution layer followed by two residual blocks. During training, we randomly crop the samples into $128 \times 128$ patches. The batch size is set to 12 and the maximum epoch of training iterations is set to 30. The Adam optimizer is used and the learning rate starts at $10^{-4}$, which decays by $50\%$ every five epochs. The number of the SGCI and SLFC blocks $D$ is set to 6; the number of dense connections in each stream of the SGCI block is set to 5; the window size and the attention head number for the STL layer are set to 8 and 8, respectively; and the size of the dynamic filter $k$ is set to 5. The codes, models, and more results will be released at: https://github.com/xufangchn/GLF-CR.

5.2. Comparisons with State-of-the-art Methods

We compare the proposed GLF-CR networks to state-of-the-art cloud removal methods, including multi-spectral based approaches, McGAN (Enomoto et al., 2017) and SpA GAN (Pan, 2020), the SAR-to-optical image translation approach, SAR2OPT (Bermudez et al., 2018), and SAR-optical data fusion based approaches, SAR-Opt-cGAN (Grohnfeldt et al., 2018) and DSen2-CR (Meraner et al., 2020). McGAN (Enomoto et al., 2017) receives the RGB channels as well as the near-infrared channel as inputs, and does full-spectrum prediction with U-Net (Ronneberger et al., 2015) as the generator. SpA GAN (Pan, 2020) takes all channels of the input optical image as input. It uses the spatial attention network (SPANet) (Wang et al., 2019) as a generator to model the map from a cloudy image to a cloudless image.
GLF-CR

Table 1
Quantitative comparisons of proposed GLF-Nets to state-of-the-art methods.

| Method           | Input               | PSNR ↑ | SSIM ↑ | SAM ↓ | MAE ↓ |
|------------------|---------------------|--------|--------|-------|-------|
| McGAN (Enomoto et al., 2017) | ✓ | 24.243 | 0.7422 | 8.8172 | 0.0543 |
| SpA GAN (Pan, 2020)        | ✓ | 26.6898 | 0.8042 | 8.8846 | 0.0372 |
| SAR2OPT (Bermudez et al., 2018) | X | 24.6623 | 0.7876 | 7.2700 | 0.0495 |
| SAR-Opt-cGAN (Grohnfeldt et al., 2018) | ✓ | 25.0971 | 0.7500 | 8.8123 | 0.0487 |
| DSen2-CR (Meraner et al., 2020) | ✓ | 31.6488 | 0.9047 | 6.3802 | 0.0215 |
| Concat (Ours)        | ✓ | 33.4809 | 0.9134 | 6.0714 | 0.0155 |
| GLF-CR (Ours)       | ✓ | 34.9317 | 0.9263 | 5.4911 | 0.0147 |

SAR2OPT (Bermudez et al., 2018) performs SAR-to-optical translation, not relying on any (cloudy) optical satellite information. Like McGAN, it also takes U-Net as the generator. SAR-Opt-cGAN (Grohnfeldt et al., 2018) and DSen2-CR (Meraner et al., 2020) both leverage the SAR image as a form of prior to guide the reconstruction process under thick, optically impenetrable clouds. The SAR’s channels are simply concatenated to the other channels of the input optical image to predict the full spectrum of optical bands. SAR-Opt-cGAN (Grohnfeldt et al., 2018) is extended from U-Net, while DSen2-CR (Meraner et al., 2020) is extended from the EDSR network (Lim et al., 2017). To validate the superiority of our method in leveraging the power of SAR images, we also refer to the fusion strategy in SAR-Opt-cGAN (Grohnfeldt et al., 2018) and DSen2-CR (Meraner et al., 2020) to train our network, by using concatenation, denoted as Concat. We concatenate the SAR’s channels and optical image’s channels as input, and remove the branch for SAR feature learning, the attention map update in the SGCI blocks, and the SLFC blocks. The quantitative results are presented in Table 1. The proposed GLF-CR network brings remarkable improvements compared to state-of-the-art methods. We choose 3 scenes to evaluate qualitative results, as shown in Fig. 5. For each scene, from top-left to bottom-right are respectively the cloudy image, the SAR image, the results from McGAN, SpA GAN, SAR2OPT, SAR-Opt-cGAN, DSen2-CR, Concat and GLF-CR, and the cloud-free image. We find that our proposed GLF-CR network achieves the best visualization performance. Detailed analyses are presented below.

We first compare the cloud removal performance of SAR-based methods to the conventional methods, McGAN and SpA GAN. As the SAR image encodes rich geometrical information about cloud-covered regions, it facilitates the ground object construction. McGAN and SpA GAN, which rely solely on cloudy optical images, are less effective than SAR-based cloud removal methods when handling images heavily covered by clouds, as shown in the second and third scenes in Fig. 5. SpA GAN performs relatively better than McGAN. SpA GAN utilizes the full-spectrum information of the cloudy optical image and is able to deal with the thin clouds. Furthermore, it uses SPANet as the generator: SPANet can collect non-local context information, and thus can promote the structural and spectral consistency of the prediction. However, both of them fail to tackle the thick cloud removal and both generate undesirable artifacts, especially for cloud-covered regions.

We next compare the cloud removal performance of the SAR-to-optical image translation approach, SAR2OPT to the SAR-to-optical data fusion based approaches. SAR2OPT, which relies solely on SAR images, can reconstruct prominent geometric characteristics related to roads, crop fields, etc. But it suffers from content vanishing because the specific potentials and peculiarities of optical images cannot be fully compensated from the SAR images. Moreover, a distinct difference in the color distribution of SAR2OPT’s reconstruction results and ground truth can be observed. SAR-Opt-cGAN (Grohnfeldt et al., 2018) adopts the same generator architecture as SAR2OPT while taking both the cloudy optical image and the SAR image as input. It performs better than SAR2OPT on PSNR but worse than SAR2OPT on SSIM. And as shown in the second scene of Fig. 5, for the road on the right, the SAR image clearly emphasizes the surface’s physical properties. SAR-Opt-cGAN fails to reconstruct it while SAR2OPT does. It demonstrates the challenge of taking advantage of multi-modal data fusion while avoiding the performance degradation caused by the undesirable effects in each modality. To some extent, DSen2-CR (Meraner et al., 2020) alleviates the performance degradation caused by the undesirable effects by utilizing a deeper generator. However, its gain is still limited.

Our methods perform favorably when compared with DSen2-CR, which exploits the inherent advantage of SAR image. Among them, Concat adopts the same approach as SAR-Opt-cGAN and DSen2-CR to utilize the complementary information embedded in SAR images. It achieves higher performance than SAR-Opt-cGAN and DSen2-CR, as shown in Table 1. Unlike the approach of SAR-Opt-cGAN and DSen2-CR, Concat contains global context interactions, which takes the information embedded in neighboring cloud-free regions into consideration, thus performing better in terms of global consistent structure. But those methods still leave distinct clouds or blur some image textures, which reflects the limitations of the concatenation method.

Furthermore, we can observe that the proposed GLF-CR network outperforms other methods by a large margin. It can restore images with more details and fewer artifacts. As shown in Fig. 5, the proposed GLF-CR network completely
restores the crop fields in the lower left corner in the first scene and the road in the second scene, while the others give over-smoothed effects. These significant improvements demonstrate that our method can better use the complementary information embedded in SAR images.
5.3. Analysis on Different Cloud Cover Levels

We further compare the proposed GLF-CR networks to state-of-the-art cloud removal methods on different cloud cover levels. We calculate the percentage of cloud cover of each image by utilizing the cloud detection flowchart in Meraner et al. (2020). Then we evaluate the performance of cloud removal on the images with cloud cover of 0% to 20%, 20% to 40%, 40% to 60%, 60% to 80%, and 80% to 100%. We show the comparison results in terms of the PSNR, SSIM, SAM, and MAE quality metrics in Fig. 6. Our methods perform favorably when compared with state-of-the-art methods on all cloud cover levels.

It is observed that the performance of multispectral-based approaches, McGAN (Enomoto et al., 2017) and SpaGAN (Pan, 2020), is negatively correlated with the cloud cover level. With the higher cloud cover level, they get less prior information and thus perform worse. And the performance of the SAR-to-optical image translation approach, SAR2OPT (Bermudez et al., 2018), is not related to the cloud cover level and relatively stable.

SAR-Opt-cGAN (Grohfeldt et al., 2018) utilizes the prior information from both cloudy images and SAR images. It suffers the performance degradation caused by the undesirable effects in both modalities. When the cloud cover is low, it is not as good as the multispectral-based methods due to the interference from additional SAR information. When the cloud cover is high, it is not as good as the SAR-to-optical image translation approach due to the interference from clouds.

DSen2-CR (Meraner et al., 2020) alleviates the performance degradation to some extent, and thus outperforms the single-modality-based methods. However, its results in terms of PSNR are not strictly negatively correlated with the cloud cover level. This indicates that DSen2-CR also suffers from the undesirable effects.

Our method considers the information embedded in neighboring cloud-free regions. Therefore, when the proportion of cloud-free regions is higher, the performance gap between our methods and DSen2-CR is larger. And the proposed GLF-CR network steadily outperforms Concat, which validates the superiority of our method in exploiting the power of SAR information.

5.4. Ablation Study

The proposed GLF-CR network improves the performance of SAR-based cloud removal by incorporating global fusion to guide the relationship among all local optical windows with SAR features and local fusion to transfer the
SAR feature corresponding to cloudy areas to compensate for the missing information. To determine what contributes to the superior performance of our proposed approach, we analyze the effectiveness of each component by comparing a few variants with and without the use of SAR image (SAR), Concatenation fusion (Concat), STL layer (STL), global fusion (GF), and dynamic filter (DF), as shown in Table 2 and Fig. 7. From the table and the figure, we can draw the following conclusions:

**Importance of SAR Image.** We validate the importance of the SAR image by training the GLF-CR network without SAR images, denoted as w/o SAR. Since the input is a single source signal, i.e., the cloudy optical image itself, a single-stream network is adopted and no fusion strategy is used. The quantitative and qualitative cloud removal results are shown in the first row of Table 2 and the fourth column of Fig. 7. Compared to the networks employing SAR images, there is a large performance gap. It tends to generate over-smoothed effects for cloud-covered regions, while the networks with SAR images can recover texture details. This demonstrates that the rich complementary information encoded in SAR images can effectively improve the cloud removal performance.

**Limitation of Concatenation Fusion.** Compared with w/o SAR, Concat only adds two channels to the input to utilize the SAR image. Its results are shown in the second row of Table 2 and the fifth column of Fig. 7. The gain of utilizing the concatenation fusion, which is 0.32dB, is limited, while our proposed GLF-CR network obtains a gain of 1.77dB, as shown in the sixth row of Table 2 and the ninth column of Fig. 7. Our proposed GLF-CR network can recover more complete texture structure and obtain better visual effects.

**Effectiveness of Global Interactions.** Capturing global interactions between contexts plays a vital role in maintaining global consistent structure. We first train the GLF-CR network by removing the STL layers in the SGCI blocks, denoted as w/o STL; the results are shown in the third row of Table 2 and the sixth column of Fig. 7. We also train the GLF-CR network by reserving the STL layer but not using the SAR feature to guide the global optical interactions, denoted as w/o GF; the results are shown in the fourth row of Table 2 and the seventh column of Fig. 7. It can be observed...
that capturing the global interactions between contexts can improve cloud removal performance effectively. It recovers clearer and more complete structure for the agricultural land in the third scene, as well as the river in the fourth scene.

**Effectiveness of SAR-Guided Global Interactions.** We further validate the effectiveness of guiding the global interactions of optical features with SAR features. Estimating the interactions from the cloudy optical image itself will introduce some error due to the observations occluded by the clouds. Since the SAR image is not affected by cloud cover, it can provide valuable guidance for capturing global interactions between contexts. We validate this point by comparing the results of \( w/o \) GF and GLF-CR. We can see that guiding the global interactions of optical features with SAR features can further improve the performance of cloud removal and make the structure of the predicted cloud-free image more consistent with ground truth.

**Effectiveness of the Dynamic Filter.** Our proposed GLF-CR network uses dynamic filtering to handle the speckle noise of SAR images. To validate the effectiveness of the dynamic filter, we train the GLF-CR network by removing the dynamic filter in SLFC blocks, denoted as \( w/o \) DF. Its results are shown in the fifth row of Table 2 and the eighth column of Fig. 7. We can see that \( w/o \) DF provides comparable results in terms of SSIM but is less effective in terms of PSNR. It generates undesirable artifacts for the sea in the first scene as well as the river in the upper left corner of the fourth scene, while our proposed GLF-CR network can alleviate the problem of speckle noise in the SAR image and generate clearer images.

6. Discussion

It is well-known that there is an inherent geometric distortion in SAR data when the terrain is undulating, due to the sensor’s sideways view. It will lead to the inconsistency between the information in the SAR data and the actual state of the ground objects, adversely affecting the cloud removal performance. The experiments in this paper are conducted on the SEN12MS-CR dataset (to our best knowledge, the only open-source cloud removal dataset with SAR data), where the SAR data is provided by the Sentinel-1 satellites. Its resolution is 10m and thus does not show excessive distortion. Furthermore, depending on the large scale of the dataset, our powerful model can address this aspect to some extent. In addition, the registration error between the optical image and its corresponding SAR image is expected to affect the learning process. The data instructions given by ESA illustrate that the Sentinel-1 SAR L1 productions and the Sentinel-2 optical L1C productions have a co-registration accuracy of within 2 pixels. We set the size of the dynamic filter in the SLFC blocks to 5 for a larger receptive field, which allows our model to work when tiny deviations exist between the SAR and optical images.

7. Conclusion

In this work, we propose a novel global-local fusion based cloud removal (GLF-CR) algorithm for high quality cloud-free image reconstruction. It boosts cloud removal performance by guiding the relationship among all local optical windows with the SAR feature and transferring the SAR feature associated with cloudy areas to compensate for missing information; further, it uses dynamic filtering to alleviate the performance degradation caused by speckle noise. Extensive experiments demonstrate that the proposed method can achieve state-of-the-art performance on a large-scale dataset.

8. Acknowledgements

F. Xu is supported by the China Scholarship Council (CSC). The work of Y. Wen is supported by the National Natural Science Foundation of China (NSFC) under Grant 61771351. The work of X. Zhu is jointly supported by the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No. [ERC-2016-StG-714087], Acronym: So2Sat), by the Helmholtz Association through the Framework of Helmholtz AI (grant number: ZT-1-PF-5-01) - Local Unit “Munich Unit @Aeronautics, Space and Transport (MASTr)” and Helmholtz Excellent Professorship “Data Science in Earth Observation - Big Data Fusion for Urban Research” (grant number: W2-W3-100), by the German Federal Ministry of Education and Research (BMBF) in the framework of the international future AI lab “AI4EO – Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond” (grant number: 01DD20001) and by German Federal Ministry of Economics and Technology in the framework of the “national center of excellence ML4Earth” (grant number: 50EE2201C).

References

Audebert, N., Le Saux, B., Lefèvre, S., 2018. Beyond rgb: Very high resolution urban remote sensing with multimodal deep networks. ISPRS Journal of Photogrammetry and Remote Sensing 140, 20–32.

Bamler, R., 2000. Principles of synthetic aperture radar. Surveys in Geophysics 21, 147–157.

Bermudez, J.D., Happ, P.N., Oliveira, D.A.B., Feitosa, R.Q., 2018. Sar to optical image synthesis for cloud removal with generative adversarial networks. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences IV-1, 5–11.

Chan, T.F., Shen, J., 2001. Non-texture inpainting by curvature-driven diffusions. Journal of Visual Communication and Image Representation 12, 436–449.

Ebel, P., Meraner, A., Schmitt, M., Zhu, X.X., 2021. Multisensor data fusion for cloud removal in global and all-season sentinel-2 imagery. IEEE Transactions on Geoscience and Remote Sensing 59, 5866–5878.

Ebel, P., Xu, Y., Schmitt, M., Zhu, X.X., 2022. Sen12ms-crp-ts: A remote-sensing data set for multimodal multitemporal cloud removal. IEEE Transactions on Geoscience and Remote Sensing 60, 1–14.

Enomoto, K., Sakurada, K., Wang, W., Fukui, H., Matsuoka, M., Nakamura, R., Kawaguchi, N., 2017. Filmy cloud removal on satellite imagery with multispectral conditional generative adversarial nets. in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 48–56.
Fang Xu, et al: Preprint submitted to Elsevier