Information Fusion 94 (2023) 215–228

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Full length article

HS^{2}P: Hierarchical spectral and structure-preserving fusion network for multimodal remote sensing image cloud and shadow removal

Yansheng Li [a], Fanyi Wei [a], Yongjun Zhang [a], Wei Chen [a], Jiayi Ma [b,*]

[a] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China
[b] Electronic Information School, Wuhan University, Wuhan 430072, China

ARTICLE INFO

Keywords:
Multimodal image fusion network
Remote sensing image cloud and shadow removal
Synthetic aperture radar-guided optical image reconstruction

ABSTRACT

Optical remote sensing images are often contaminated by clouds and shadows, resulting in missing data, which greatly hinders consistent Earth observation missions. Cloud and shadow removal is one of the most important tasks in optical remote sensing image processing. Due to the characteristics of active imaging that enable synthetic aperture radar (SAR) to penetrate cloud cover and other climatic conditions, SAR data are extensively utilized to guide optical remote sensing image cloud and shadow removal. Nevertheless, SAR data are highly corrupted by speckle noise, which generates artifact pollution to spectral features extracted from optical images and makes SAR-optical fusion ill-posed to generate cloud and shadow removal results while retaining high spectral fidelity and reasonable spatial structures. To overcome the aforementioned drawbacks, this paper presents a novel hierarchical spectral and structure-preserving fusion network (HS^{2}P), which can recover cloud and shadow regions in optical remote sensing imagery based on the hierarchical fusion of optical and SAR remote sensing imagery. In HS^{2}P, we present a deep hierarchical architecture with stacked residual groups (ResGroups), which progressively constrains the reconstruction. To pursue the adaptive selection of more informative features for fusion and reduce attention to the features with artifacts brought by clouds and shadows in optical data or speckle noise in SAR data, residual blocks with a channel attention mechanism (RBCA) are recommended. Additionally, a novel collaborative optimization loss function is proposed to preserve spectral features while enhancing structural details. Extensive experiments on the publicly open dataset (*i.e.*, SEN12MS-CR) demonstrate that the proposed method can robustly recover diverse ground information in optical remote sensing imagery with various cloud types. Compared with the state-of-the-art cloud and shadow removal methods, our HS^{2}P achieves significant improvements in terms of quantitative and qualitative results. The source code is publicly available at https://github.com/weifanyi515/HS2P.

1. Introduction

Continuous monitoring of Earth's surface has a vital role in understanding the world [1]. With the rapid growth of remote sensing technology, optical remote sensing images have gradually become the mainstream way to monitor Earth's surface. However, optical remote sensing images are unavoidably contaminated by clouds, leading to noncontinuous observations of Earth's surface. According to the analysis of USGS data, the average global annual cloud coverage is approximately 66% [2]. And the statistics of Landsat ETM+ data reveal that 35% of land areas are approximately covered by clouds [3]. Therefore, cloud cover substantially hinders the wide application of optical remote sensing images, as clouds in optical remote sensing images tremendously affect various Earth monitoring tasks, which involve seamless continuous observations. To assure the seamless observation of Earth's surface, cloud and shadow removal in optical remote sensing imagery has become an urgent problem.

Generally, cloud and shadow removal in optical remote sensing imagery is aimed at reconstructing the missed remote sensing image data contaminated by clouds by leveraging the complementary information. According to the difference in the auxiliary information type, cloud and shadow removal approaches can be categorized into three major clusters: single-image reconstruction approaches, multitemporal fusion approaches and multimodal fusion approaches. Single-image reconstruction approaches fill in the missing data regions with original scene information from the remaining spatial parts or other spectra [4–7]. These approaches assume that auxiliary clean spatial regions or spectra exist and usually fail to reconstruct large or thick opaque

* Corresponding author.
E-mail addresses: yansheng.li@whu.edu.cn (Y. Li), weifanyi@whu.edu.cn (F. Wei), zhangyj@whu.edu.cn (Y. Zhang), weichenrs@whu.edu.cn (W. Chen), jyma2010@gmail.com (J. Ma).

https://doi.org/10.1016/j.inffus.2023.02.002
Received 25 October 2022; Received in revised form 30 January 2023; Accepted 1 February 2023
Available online 4 February 2023

cloud-covered scenes. Multitemporal approaches use the same scenes from other periods to recover the missing ground information [8–11]. The limitation of multitemporal approaches is the assumption that there is a slight difference between the data acquired at different periods. However, the temporal stability of land cover cannot be ensured, which renders multitemporal cloud removal results to serve fine-grained monitoring or change detection approaches [12]. In the case of multimodal fusion approaches, cloud removal is supported by an additional data source [13,14]. Different sensors have diverse imaging principles, and the focus in describing the scene of the multimodal images captured by them is significantly different [15–17]. Therefore, multimodal fusion approaches can reconstruct obscured regions by fusing complementary information in different modal images, which exhibits excellent potential. One of the most concerning topics is the fusion of synthetic aperture radar (SAR) data and optical data. SAR is an all-weather sensor that records the intensity of the radar backscattering. SAR is also capable of collecting ground information regardless of clouds due to the advantage of strong penetrability, which offers complementary contextual and structural information to adequately compensate for the contaminated regions in optical images [18]. Based on this property, cloud and shadow removal in optical remote sensing imagery to exploit SAR data is considered in this paper.

In recent years, cloud and shadow removal methods based on the fusion of SAR data and optical data have showed strong performance. Nevertheless, SAR data are highly corrupted by speckle noise due to coherent processing of backscattered signals [19], which brings artifact pollution to spectral features extracted from the input optical images and makes SAR-optical fusion ill-posed to generate cloud removal results while retaining high spectral fidelity and reasonable spatial structures. Especially, the reconstruction of small or thin cloud-covered optical images that provide a large amount of uncontaminated spectral information is susceptible to speckle noise, which leads to the generation of cloud removal results with fuzzy details. Although SAR-optical fusion-based cloud and shadow removal methods have been improved over the years, a majority of the existing methods ignored the undesirable effect brought by speckle noise when utilizing SAR data as auxiliary input information and directly stacked SAR data and optical data for fusion [13,20,21]. In order to reduce attention to channelwise features with artifacts produced by clouds and shadows in optical images or speckle noise in SAR images while emphasizing more informative features adaptively in SAR-optical fusion, we use residual blocks with a channel attention mechanism (RBCA) to form the deep network for cloud and shadow removal in this paper. Furthermore, a few SAR-optical fusion-based methods focused on reconstructing spectral information similar to the specified targets by elementwise losses, while ignoring geometric structural information in cloud and shadow removal results [21–23]. To tackle this limitation, we design a collaborative optimization loss function that contains a spectral preserving loss and structural preserving loss to operate our network to reconstruct rich spectral and structural information.

With the aforementioned considerations, this paper proposes a hierarchical spectral and structure-preserving fusion network (HS^2P), which can reconstruct cloud and shadow regions based on the fusion of optical data and SAR data. The architecture of HS^2P is designed to progressively constrain the reconstruction with the stacked residual groups (ResGroups) to guarantee the quality of cloud removal results on multiple levels of the deep network. This architecture is also beneficial to shallow feature delivery. To reduce artifacts in the cloud removal results, we exploit RBCA as basic components of ResGroups to guide the network to adaptively select more informative channelwise features for fusion. In a further step, a novel collaborative optimization loss function is developed to preserve spectral features while enhancing structural details in cloud removal results. We conduct experiments on the public large-scale dataset (*i.e.*, SEN12MS-CR). The experimental results show that our proposed method reconstructs diverse ground information with higher spectral fidelity and richer structural textures in optical remote sensing imagery covered by various types of clouds. Our method is also superior to the state-of-the-art cloud and shadow removal methods in both quantitative evaluations and qualitative evaluations.

Overall, the main contributions of this paper are summarized as follows:

- This paper proposes a hierarchical spectral and structure-preserving fusion network named HS^2P, which progressively reconstructs cloud and shadow regions.
- In HS^2P, residual blocks with a channel attention mechanism named RBCA are exploited as basic components. The embedded attention module guides the network to emphasize more informative features of multimodal imagery.
- We introduce a collaborative optimization loss, which enables our HS^2P to learn more powerful spectral and structural feature representations and to enhance spectral fidelity and prominent structural features in cloud removal results.

The remainder of this paper is organized as follows: Section 2 briefly reviews the related work. In Section 3, we introduce our proposed method in detail. Section 4 provides the experimental results, followed by a discussion of the experimental critical parameters. The conclusion is given in Section 5.

2. Related work

In this section, we review the related cloud and shadow removal methods via deep learning. Generative adversarial learning-driven methods and residual learning-driven methods are introduced in Sections 2.1 and 2.2, respectively. And Section 2.3 introduces some advanced methods that are embedded with attention mechanisms.

2.1. Generative adversarial learning-driven cloud and shadow removal

With the maturity of deep learning, approaches for cloud and shadow removal have been constantly developed. Generative adversarial networks (GANs) have experienced a massive rise in popularity among deep learning-based methods. A GAN consists of a generator and discriminator. The goal of the generator is to yield images that the discriminator cannot recognize, and the discriminator's goal is to distinguish between actual images and generated images as accurately as possible. The generator-discriminator game makes the GAN generate images that are similar to the corresponding targets. Bermudez et al. proposed a method to map cloud-free optical images from co-registered SAR images based on the image translation capability of a conditional generative adversarial network (cGAN), which reconstructed scenes depending only on SAR data [24]. However, since SAR data lack information in spectral aspect, it is hard to transform SAR images to cloud-free multispectral images in good quality. In order to address the above problem, methods that eliminated clouds and shadows by synergistically utilizing the complementarity of SAR data and optical data were proposed. Grohnfeldt et al. developed a SAR-Optical-cGAN based on the Pix2Pix model [25] and removed synthetic clouds with SAR data fusion [13]. Gao et al. further considered the weak correlation between SAR data and optical data, which converted SAR images into simulated optical images with strong complementarity first and generated cloud-free images using both simulated optical images, SAR images and cloudy optical images [14]. Following this idea, Gao et al. advanced to balance the global loss, local loss, perceptual loss and GAN loss in their work [26]. The local loss makes the network pay more attention on the reconstruction of missing regions and the perceptual loss leads to results with better visual perception. For presenting more promising cloud and shadow removal results, a spatiotemporal generator network (STGAN) was proposed, which added addition multitemporal

information as input [27]. However, it traded temporal resolution, thus reducing the possibility of seamless monitoring. Subsequently, Darbaghshahi et al. proposed two-stage GANs used for SAR to optical translation and cloud removal respectively and improved vanilla U-net architecture by utilizing dilated convolutions to increase receptive view and prevent missing information, which made progress on removing clouds in optical images consist of four bands (RGB and NIR) [20]. Furthermore, cycle-GANs were also applied for reducing dependence on paired cloudy and cloud-free training data. A cycle-consistent GAN was exploited for unpaired image translation [28]. Nevertheless, the prevalent problem for GANs is the tendency to generate fake details or unexpected artifacts because the generator and discriminator have difficulty achieving the theoretical Nash equilibrium in the training process [29,30].

2.2. Residual learning-driven cloud and shadow removal

A deep residual network (ResNet) [31] exploits residual blocks (ResBlocks) as basic components. Each ResBlock is composed of several layers, and its output is the sum of its last layer and its input. In this way, the layers within the ResBlock are forced to learn the difference between input and output, which usually corresponds to noise corruption in a noisy image [22]. Residual learning can also quickly optimize large and deep networks and stabilize performance [32]. It has been reported that many vision tasks can be further improved by simply replacing plain convolutional neural networks (CNNs) with ResNets [33]. Hence, ResNets are utilized frequently to reconstruct the contaminated areas in cloud and shadow removal tasks. Li et al. introduced a deep residual symmetrical concatenation network (RSC-Net), which was designed as a symmetrical architecture consisting of multiple residual convolutional layers and residual deconvolutional layers [22]. The cloud-free details can be passed to the top layers directly by symmetrical concatenations between the convolutional layers and deconvolutional layers, thus alleviating the damage to the input cloud-free regions. Meraner et al. implemented the similar idea by employing a long skip connection in their DSen2-CR [21]. DSen2-CR not only handled the presence of thin clouds, but also achieved superiority of cloud removal for heavily occluded images. In addition to the conceptual considerations, a used large dataset is also needed for promising the generalization capability of the networks. For this, Meraner et al. released the globally sampled SEN12MS-CR dataset containing triplets of cloudy Sentinel-2 optical images, cloud-free Sentinel-2 optical images and Sentinel-1 SAR images, which promoted cloud removal researches based on SAR-optical fusion. However, non-local features cannot be effectively represented in the DSen2-CR model. To solve this problem, a multiscale deep ResNet (MDRN) with the embedding of multiscale convolution units was proposed [23]. Profitting from these units, the MDRN has larger receptive fields to extract multiscale features. He et al. proposed a deformable context feature pyramid (DCFP) module, which replaced fixed filter receptive fields to an adaptive manner based on the shapes and sizes of the clouds [34]. However, ResNets easily yield unsatisfactory cloud and shadow removal results when handling complex scenes [35], which deserves further improvement.

2.3. Attention mechanisms

Aimed at enhancing the representativeness of the extracted features, using attention mechanisms is another choice. The attention mechanisms are beneficial for image information reconstruction by guiding the available processing resources to the most informative input components [36]. A network composed of several dense spatial attention blocks (DSAB) was designed [37]. The basic component of DSAB is the convolution block attention module (CBAM), which contains a channel attention module and spatial attention module [38]. The feature maps of the intermediate layers are refined in the channel and spatial dimensions to different degrees by the two sequential submodules. Zhou et al. further integrated both channel attention blocks and multiscale convolution blocks [39] in their multiple scale attention ResNet (MSAR-Defognet). Moreover, the channel attention mechanism was proved to be effective for restoring thin cloud-covered scenes [40]. In order to pay more attention to the recovery of cloudy areas, Xu et al. designed an attention module which is able to generate attention maps optimized by cloud masks in their attention mechanism-based GAN (AMGAN-CR) [35]. Extended from graph neural networks (GNNs), a spatiotemporal reasoning module (STeRe) was proposed to construct long-range dependencies through a differentiable attention mechanism while preserving spatial information of nodes in the graph, thus tracking blur or dense target objects effectively [41]. Recently, He et al. advanced an idea of employing the transformer to capture long-range dependencies between multimodal data and proposed an attentive information aggregation mechanism to aggregate heterogeneous information based on the self-attention mechanism [42]. Take advantages of the transformer, Xu et al. presented a SAR-guided global context interaction (SGCI) block in their global–local fusion-based cloud removal method (GLF-CR). The SGCI block guides the reconstructed regions to maintain consistent structure with cloud-free regions by SAR features [43].

3. Methodology

In this section, an introduction of our proposed HS^{2}P is presented. Section 3.1 overviews the proposed approach. Then, Section 3.2 adequately introduces the deep hierarchical architecture. Next, we present the detailed introduction of RBCA in Section 3.3. Moreover, the custom loss is given in Section 3.4.

3.1. Overview of the proposed approach

To pursue the accurate reconstruction of regions covered by clouds and shadows, we develop HS^{2}P based on SAR-optical fusion, as shown in Fig. 1. A data fusion module is employed in HS^{2}P, which has a concatenation layer followed by a convolution layer and an attention module to fuse the input paired SAR image d_{SAR} and optical image d_{OPT}. In HS^{2}P, there are N stacked ResGroups that construct the deep hierarchical architecture. The ResGroups at multiple levels of the HS^{2}P generate hierarchical outputs in both spectral term and structural term during the training phase. For the interior of ResGroups, RBCA are the basic components.

To reconstruct both spectral features and structural features with high quality, we propose a new collaborative optimization loss function to optimize our HS^{2}P. Note that our approach adaptively reconstructs the blocked regions without relying on accurate cloud and shadow detection results. In the cloud and shadow removal domain, some approaches utilize cloud masks to divide cloudy areas and clear areas in optical remote sensing images [44,45]. Then, the masked regions were regarded as blank regions for information reconstruction. However, the visibility of ground covered by different thicknesses of clouds varies. Thin and translucent clouds obscure only the spectral information and abundant features can still be extracted from scenes in this circumstance, while optically thick clouds completely occlude the ground, causing nearly all the ground information to be lost. Consequently, it is better not to treat all cloudy areas as blank regions. The detailed introduction of HS^{2}P is described as follows:

3.2. Deep hierarchical architecture

As Fig. 1 shows, the stacked ResGroups form the trunk part of HS^{2}P. After N ResGroups, the final output of HS^{2}P is obtained. The information flow of the ResGroups is formulated as:

$$F_g = RG_g\left(F_{g-1}\right) = RG_g\left(RG_{g-1}\left(\ldots RG_1\left(F_0\right)\right)\right), \quad (1)$$

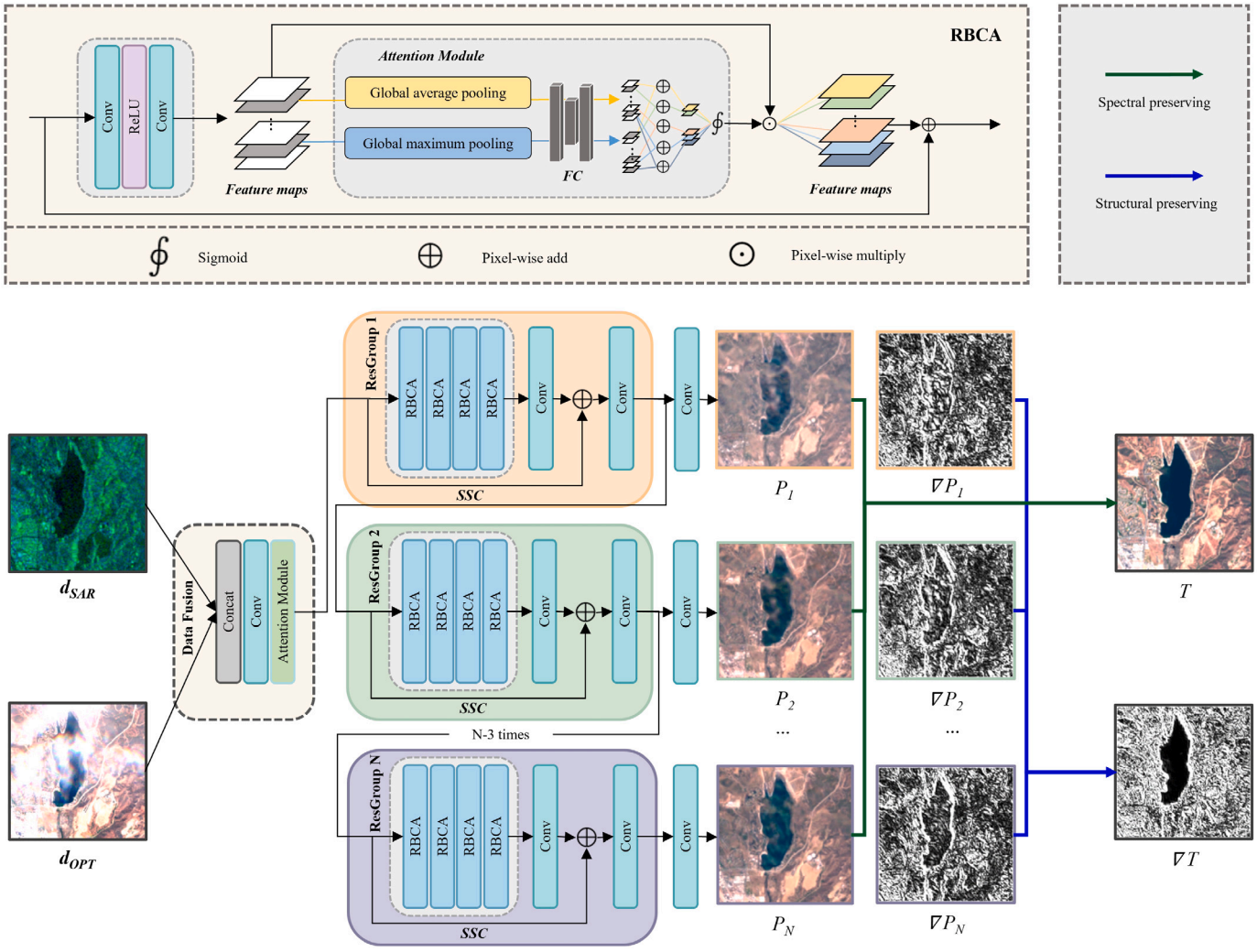

Fig. 1. Architecture of HS^{2}P. The top rectangle represents RBCA. The black arrow represents the information flow, the green arrow represents the spectral preserving function, and the blue arrow represents the structural preserving function.

where RG_g represents the gth ResGroup, while $g = 1, 2 \ldots N$ and N is the total number of ResGroups in HS^{2}P. Feature maps generated by the gth ResGroup are denoted by $F_g \epsilon R^{f_c \times W \times H}$, where we discard the batch dimension from our notations. f_c represents the number of feature maps in F_g, and W and H represent the width and height, respectively, of a feature map. It is remarkable that $F_0 \epsilon R^{f_c \times W \times H}$ is the output of the data fusion module, which is formulated by Eq. (2).

$$F_0 = DF\left(d_{SAR}, d_{OPT}\right), \tag{2}$$

where DF represents the function of the data fusion module. $d_{SAR} \epsilon R^{f_s \times W \times H}$ and $d_{OPT} \epsilon R^{f_o \times W \times H}$ represent the input SAR data and input optical data, respectively. Note that $f_s = 2$ and $f_o = 13$, which is consistent with the original data bands. The ResGroups generate hierarchical outputs in spectral and structural terms. Inspired by the previous work [46], we use a short skip connection (SSC) in ResGroups to make it learn information at a coarse level and to stabilize the training process. Then, a convolutional layer is applied after each ResGroup to permute the dimensions of F_g to its original format $\left(R^{f_c \times W \times H} \longrightarrow R^{f_o \times W \times H}\right)$. There are N clear images $P_n (1 \leqslant n \leqslant N)$ generated for each input d_{SAR} and d_{OPT} in the training process, which are further utilized for loss computation with the corresponding real, clear optical image T. Auxiliary gradient information has an important role in alleviating blurry geometric structures [47]. Therefore, we obtain N gradient maps $\nabla P_n (1 \leqslant n \leqslant N)$ from the N generated clear images to represent textual features by computing the difference between adjacent elements in an image. Then, we utilize them for loss computation with the real gradient map ∇T, which is extracted from the corresponding T. The hierarchical outputs of our HS^{2}P are formulated by Eq. (3).

$$\begin{aligned} P(d) = &\{\{P_1(d), P_2(d), \ldots, P_N(d)\}, \\ &\{\nabla P_1(d), \nabla P_2(d), \ldots, \nabla P_N(d)\}\}, \end{aligned} \tag{3}$$

where $d = [d_{SAR}, d_{OPT}]$ represents the input. P_1 to P_N are the N generated clear images in spectral space, and ∇P_1 to ∇P_N are the N gradient maps in structural space. The hierarchical outputs are applied to constrain the reconstruction of the cloud and shadow regions at multiple levels of the network, which makes the restored information repetitiously refined to avoid prominent distortion and to enhance the quality and fidelity of the generated cloud-free results.

3.3. Residual block with channel attention mechanism

Hu et al. proposed an attention mechanism and confirmed that the attention mechanism allows the network to perform feature recalibration, through which it can learn to use global information to selectively emphasize informative features and suppress less useful features [36]. Inspired by this work, we exploit ResBlocks embedded with a channel attention mechanism, which is named RBCA, to form ResGroups in the proposed network. In RBCA, we employ a convolutional layer followed by a ReLU layer and another convolutional layer to extract multimodal

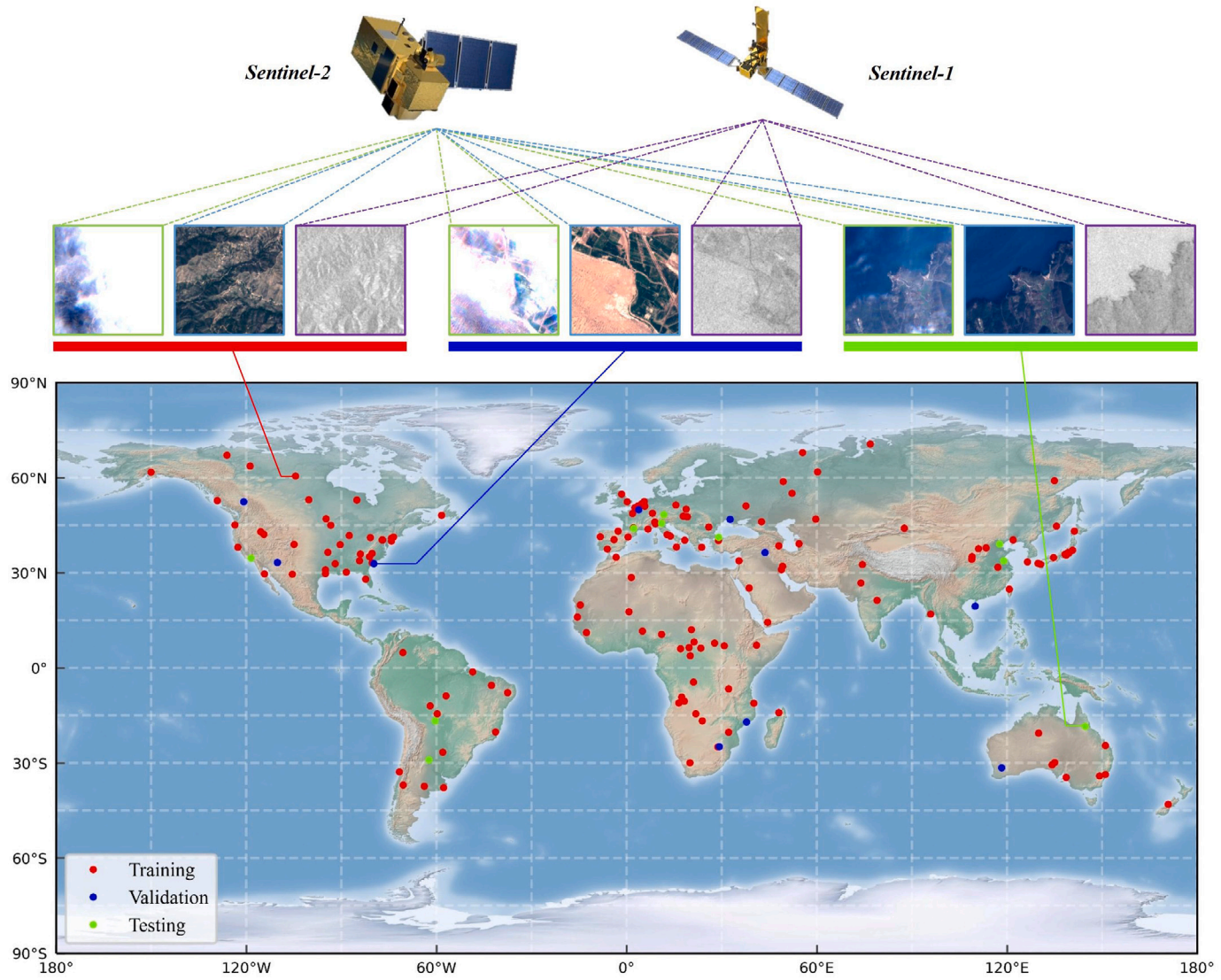

Fig. 2. Distributions of the training set, validation set and testing set. Red points represent the training set, Blue points represent the validation set, and Green points represent the testing set. Some triplets are shown above the map.

features. A global average pooling layer and global maximum pooling layer are then utilized in parallel to compress the extracted features into channel descriptors, which balances both the mean and extreme standards. The channel descriptors are then fed into two fully connected (FC) layers, achieving the capture of channelwise dependencies in a flexible and nonmutually exclusive way of learning. The average channel descriptors and maximum channel descriptors produced by FC are added, which will be further used to generate the final weight of each channel by a sigmoid layer. We then adjust the feature maps by the computed weights using multiplication. The last step in RBCA is adding the input and adjusted feature maps to execute residual learning. RBCA is defined as follows:

$$F_b(d) = F_{b-1}(d) + W_{b-1}(d) \times Conv(ReLU(Conv(F_{b-1}(d)))), \tag{4}$$

where F_b is the output of the bth RBCA and the input of the b+1th RBCA. Similarly, F_{b-1} represents the input of the bth RBCA. The produced channelwise weights in the bth RBCA are denoted by W_{b-1}. $Conv$ and $ReLU$ denote the convolution layer and ReLU layer, respectively. With the embedding of the channel attention mechanism, RBCA can adaptively recalibrate multimodal features in a channelwise manner and facilitate the quality of the network representations. The same channel attention mechanism is also applied in the data fusion module, which is displayed in Fig. 1. After the data fusion module, multimodal features are extracted and discriminatively merged instead of being simply concatenated channelwise.

3.4. Collaborative optimization loss

Generally, existing methods of cloud and shadow removal use $\mathcal{L}_1$ loss for information reconstruction, which disregards the structural information. Inspired by the previous work [48], we design a collaborative optimization loss to retain spectral and structural information, which is composed of a spectral preserving loss and structural preserving loss. The custom loss is defined as:

$$\begin{aligned}\mathcal{L}_{S^2P} &= \sum_{n=1}^{N} \lambda_n (\mathcal{L}_{SP}^n + \alpha \mathcal{L}_{ST}^n) \\ &= \sum_{n=1}^{N} \lambda_n (\|P_n - T\|_1 + \alpha \|\nabla P_n - \nabla T\|_1),\end{aligned} \tag{5}$$

where $\mathcal{L}_{SP}^n$ and $\mathcal{L}_{ST}^n$ denote the spectral preserving loss and structural preserving loss, respectively, of the nth ($1 \leqslant n \leqslant N$) ResGroup. P_n and ∇P_n are the generated clear image and gradient map, respectively, of

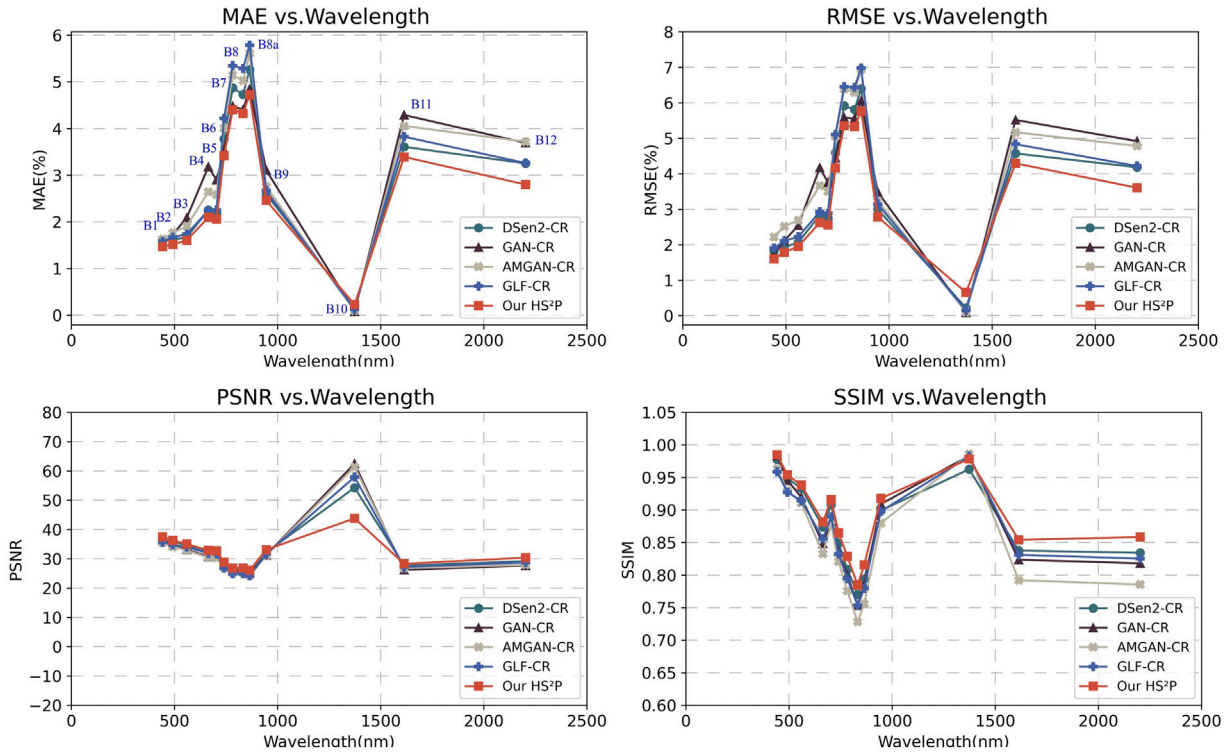

Fig. 3. Variation in average metrics with spectral bands.

the nth ResGroup. α is a regularization constant to adjust the weight of the two losses. λ_n is another regularization constant, which represents the weight for different ResGroups.

Natekin et al. indicated that the $\mathcal{L}_1$ loss can provide robustness to outliers [49]. Therefore, we use the $\mathcal{L}_1$ loss as the spectral preserving loss $\mathcal{L}_{SP}$ and structural preserving loss $\mathcal{L}_{ST}$. The underlying idea of $\mathcal{L}_{SP}$ is to enhance the similarity in the spectral term between the predicted images and the target images. The gradient maps represent the gradient lengths, considered the gradient intensity, which are adequate to reveal the sharpness of local regions in a given image. By calculating the $\mathcal{L}_1$ loss between the two given gradient maps ∇P_n and ∇T, $\mathcal{L}_{ST}$ helps the model learn from the gradient space and capture the structure dependency, which makes the predicted cloud-free images have textures similar to the targets. With our custom loss, the network is optimized in both spectral aspects and structural aspects to generate cloud removal results with not only a fine appearance but also explicit outlines. In addition, it is commonly recognized that the deeper layers of the network are accompanied by stronger nonlinear representations. In this regard, the value of λ_n is designed to increase as n increases, which means that the outputs of deeper ResGroups are allocated to larger weights. In practice, we design a monotone increasing function for λ_n utilizing the sigmoid function.

4. Experimental results

Section 4.1 introduces the experimental data, including the description and distribution of the adopted dataset SEN12MS-CR. The metrics for quantitative evaluations are also introduced in Section 4.1. The experimental setup of this paper is given in Section 4.2 in detail. From the prediction perspective, Section 4.3 quantitatively and qualitatively reports the cloud removal results of our method and the state-of-the-art cloud and shadow removal methods. Section 4.4 further explains our ablation study to confirm the effectiveness of our contributions. Next, we analyze the experimental critical parameters in Section 4.5, including the ResGroups number and the weight of our collaborative optimization loss. Finally, Section 4.6 presents the application of the proposed HS^{2}P on large-scale scenes.

4.1. Experimental data and evaluation metrics

4.1.1. Dataset description

To fully show the effectiveness of the presented method, we conduct experiments on the public large-scale dataset named SEN12MS-CR [50], which contains triplets of cloudy Sentinel-2 optical images, cloud-free Sentinel-2 optical images and Sentinel-1 SAR images. The publicly released SEN12MS-CR contains 175 nonoverlapping regions of interest (ROIs), each of which is cut into several small patches with a size of 256 × 256 pixels and strides of 128 pixels. There are 122,218 triplets with 10 m spatial resolution in total. These patches are sampled over Earth's land mass and four meteorological seasons. We divide all the patches into three subdatasets according to ROIs, namely, the training set, validation set and testing set. We also ensure that every subdataset is distributed across the four meteorological seasons. The distributions of our three subdatasets are shown in Fig. 2, where red points symbolize the training set, blue points symbolize the validation set and green points mark the locations of the testing set. In particular, such a division implements regional nonoverlapping among the three subdatasets, which guarantees the global universality of our approach.

In SEN12MS-CR, Level-1C top-of-atmosphere reflectance products are selected as Sentinel-2 data. For Sentinel-1 data, the Sentinel-1 IW Level-1 GRD products are chosen, and the values are backscatter coefficients that have been transformed into dB scales. To reduce the temporal difference that may be caused by building changes or vegetation, all triplets from the same scenes are guaranteed to be acquired within the same meteorological season. In our experiments, both polarization channels (VV and VH) in Sentinel-1 SAR data are utilized. To fully exploit the spectra, we use all 13 bands (B1, B2, B3, B4, B5, B6, B7, B8, B8a, B9, B10, B11, and B12) in the Sentinel-2 optical data.

4.1.2. Quantitative evaluation metrics

In our experiments, we utilize several common metrics to quantitatively evaluate the performance of our proposed model. These metrics include the mean absolute error (MAE), root-mean-square

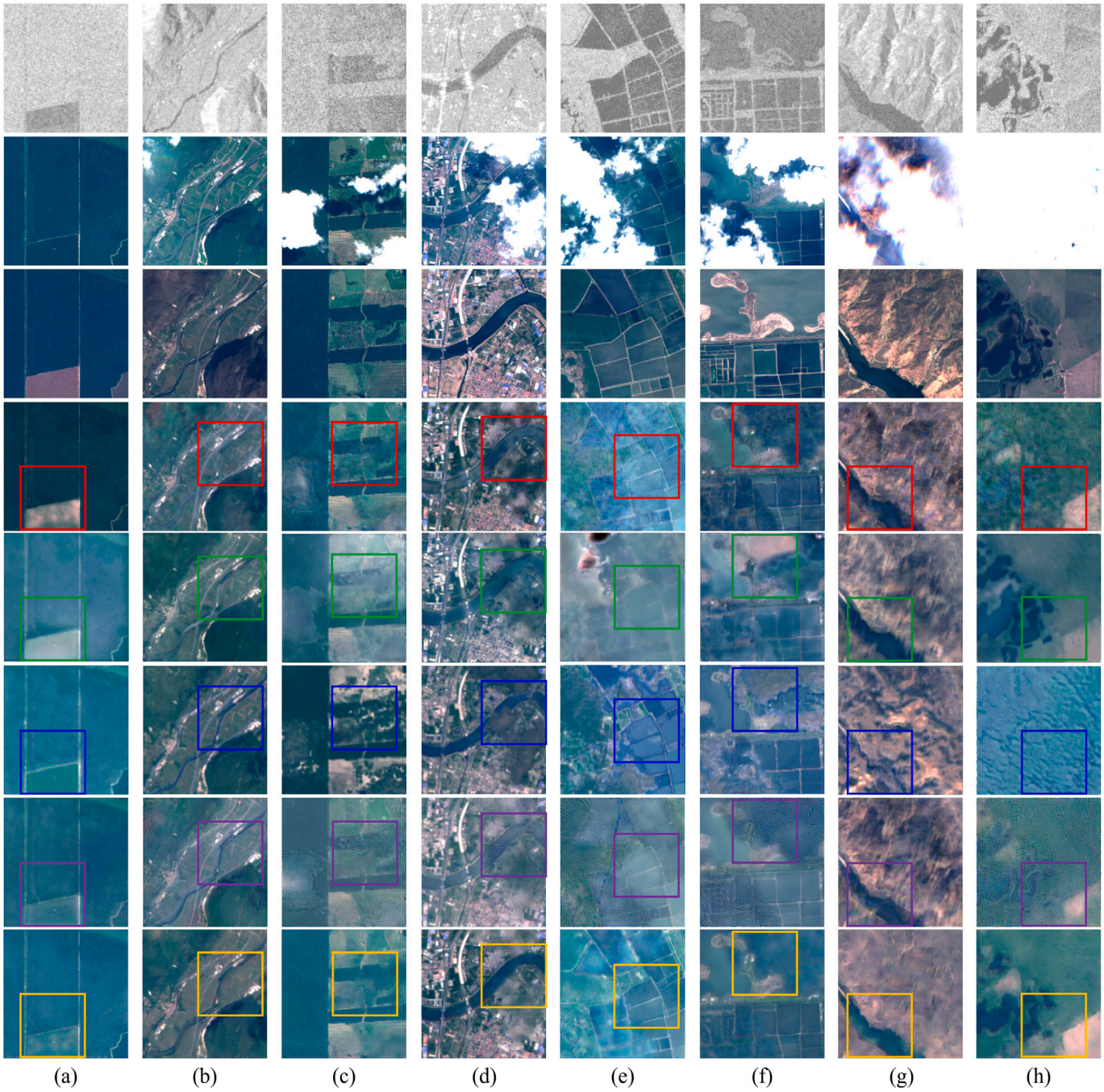

Fig. 4. Cloud and shadow removal results classified by cloud types. The first row shows the Sentinel-1 SAR images, the second row shows the cloudy Sentinel-2 optical images, the third row shows the cloud-free Sentinel-2 optical images, the fourth row shows the results of the DSen2-CR model, the fifth row shows the results of the GAN-CR model, the sixth row shows the results of the AMGAN-CR model, the seventh row shows the results of the GLF-CR model and the last row shows the results of the HS^{2}P model.

error (RMSE), peak signal-to-noise ratio (PSNR), spectral angle mapper (SAM) [51] and structural similarity (SSIM) [52]. Both MAE and RMSE are commonly utilized elementwise error indicators. Distinctively, RMSE calculates the square of the error first, which magnifies large deviation. The lower values of MAE and RMSE indicate the higher precision of the evaluated images. The PSNR is another elementwise metric used to assess the quality of recovered images. The PSNR values are proportional to the quality of the predictions. The SAM is an imagewise metric that treats spectra as high-dimensional vectors and quantifies the similarity between two given images by calculating the angle between the vectors. The lower the values of SAM are, the higher the similarity between targets and predictions. The SSIM is also imagewise and is designed to capture structural similarity by means of quantifying differences between two images in terms of luminance, contrast and structure. The SSIM values are within the range of $[0, 1]$, positively correlating to the structural quality of the predicted images. For multispectral images, SSIM is calculated by taking the average of separate calculations for each band.

4.2. Experimental setup

The proposed model is trained within 30 epochs on a NVIDI GeForce GTX 1080 Ti. For data preparation, we create patches with the operations of sort shuffling, size clipping (128 × 128 pixels), random rotations and flipping. The input optical data are clipped into the range of $[0, 10,000]$, while the clipping ranges are $[-25, 0]$ and $[-32.5, 0]$ for

Table 1
Average quantitative results of different methods.

Method	MAE (↓)	RMSE (↓)	PSNR (↑)	SAM (↓)	SSIM (↑)
DSen2-CR [21]	0.0289	0.0412	28.3537	8.5748	0.8766
GAN-CR [20]	0.0306	0.0447	28.0835	9.9437	0.8667
AMGAN-CR [35]	0.0315	0.0464	27.4809	9.9841	0.8490
GLF-CR [43]	0.0308	0.0446	27.7344	10.9992	0.8648
Our HS^{2}P	**0.0265**	**0.0376**	**29.3851**	**7.8649**	**0.8896**

Table 2
Ablation study on different modules in HS^{2}P.

Module			MAE (↓)	RMSE (↓)	PSNR (↑)	SAM (↓)	SSIM (↑)
RBCA	HSP	HST					
✘	✘	✘	0.0299	0.0443	27.9997	8.5905	0.8770
✔	✘	✘	0.0307	0.0433	28.1990	8.8448	0.8748
✔	✔	✘	0.0289	0.0402	28.6080	8.4505	0.8840
✔	✔	✔	**0.0265**	**0.0376**	**29.3851**	**7.8649**	**0.8896**

VV polarizations and VH polarizations, respectively, of SAR data [21]. In addition, all the bands of Sentinel-2 data are divided by 2000 to guarantee numerical stability [53]. Analogously, a scaling operation is applied to Sentinel-1 SAR data to match the distribution of optical data [21]. During the training process, we use the validation set to evaluate the trained model every epoch and select the model with the best performance. Since PSNR and SSIM are the most extensively utilized image objective evaluation metrics, the best performing model is determined by using them to evaluate the validation set after each epoch in the training phase. Specifically, the model with the highest PS (*i.e.*, $PS = PSNR + SSIM \times 10.0$) on the validation set will be selected for testing. In the testing phase, the predictions are compared with the targets to assess the performance of the trained model. To visualize the cloud removal results, the predicted images are multiplied by 2000 to revert to their original range $[0, 10,000]$. Then, we synthesize the RGB image for a given Sentinel-2 image by concatenating its B4, B3, and B2 spectra. Due to the polarization imaging mode of Sentinel-1 SAR data, grayscale images of the single VV band are used for demonstration.

4.3. Comparison with the state-of-the-art methods

As representative methods of cloud and shadow removal, DSen2-CR [21], GAN-CR [20], AMGAN-CR [35] and GLF-CR [43] are selected for comparison in our experiments. Specifically, Table 1 presents the average quantitative experimental results of different methods. The optimal values are marked in bold, similar to the following table.

The average quantitative results show that our HS^{2}P can obtain the optimal values on all the metrics, which proves the superiority of our method in a straightforward way. In particular, our method makes significant promotion on MAE/RMSE/SAM by ∼8/9/8% compared with DSen2-CR. For comprehensively evaluating the reconstruction of each band in multispectral images, we compare the variation in the average metrics with spectral bands as displayed in Fig. 3. Due to SAM is a metric for multispectral data, it is not compared in Fig. 3 by band. It is seen that the proposed method prominently performs in the vast majority of comparisons, and only relatively poor results are observed in B10.

Figs. 4 and 5 present the cloud and shadow removal results of the DSen2-CR model, GAN-CR model, AMGAN-CR model, GLF-CR model and the proposed model. Fig. 4 shows the results classified by cloud types, in which Columns a and b are almost clear scenes, Columns $c - f$ are small cloud-covered scenes, and Columns g and h are large cloud-covered scenes. The proposed method can handle various types of clouds and reconstruct significant features even when clouds almost completely block the ground. In general, the results of our model show clear borders and sharp textures, while the other comparative models tend to generate more ambiguous results which can be observed obviously in Columns e and f of Fig. 4. Fig. 5 shows the results classified by land cover, in which Columns a and b are mountains, Columns c and d are waters, Columns e and f are croplands, and Columns g and h are urban areas. The comparative models easily generate results with fuzzy ground objects. There are even artifacts left in some scenes, as Columns c, d, g, and h of Fig. 5 display. Due to the AMGAN-CR model relies solely on cloudy optical images, it suffers more from the generated fuzzy features than the other multimodal cloud removal methods as Row 6 of Fig. 5 shows. Furthermore, Columns f and g of Fig. 5 prominently demonstrate that HS^{2}P is superior in continuous features generation, which means that HS^{2}P tends to recover complete ground objects corrupted by clouds. As our proposed model utilizes RBCA to adaptively select more informative features instead of concatenating them directly in the channel dimension compared to DSen2-CR and considers the enhancement on the reconstruction of structural details compared to GAN-CR and GLF-CR, it alleviates undesirable artifacts that break consistent structure of the cloud removal results, thus significantly outperforming the comparative models across a variety of challenging terrains (*i.e.*, mountains with complex textures and urban areas with numerous ground objects) and visually demonstrating good results.

The absolute error maps of the cloud and shadow removal results containing different cloud types and land cover, which are shown in Fig. 6, verify the effectiveness of our method in a further step. Rows 2–6 of Fig. 6 display the cloud and shadow removal results of the DSen2-CR model, GAN-CR model, AMGAN-CR model, GLF-CR model and the proposed model with their absolute error maps. We observe from the selected scenes that the proposed model largely retains the proper features of the ground truth. Generally, our method generates lower error in detail, but the other methods are more prone to generate artifacts resulting in poor perceptual quality. Overall, the comparative experimental results displayed above indicate that our method surpasses the other selected state-of-the-art methods with several quantitative and qualitative contrasting approaches, which confirms the effectiveness and superiority of the proposed HS^{2}P.

4.4. Ablation study of the presented HS^{2}P

To further demonstrate the effectiveness and necessity of our contributions, we conduct an ablation study on the proposed model. The influence of our contributions is explained here.

To show the effect of the applied channel attention mechanism in RBCA, we obtain feature maps before and after the attention module of the last RBCA in HS^{2}P and convert them to heatmaps. There are not enough clear boundaries and outlines extracted before the channel attention module, which is shown on the second row of Fig. 7. However, important features that are similar to the corresponding target are emphasized by the attentional module, as the third row of Fig. 7 shows. Note that the regions in red boxes show relatively blurry features, and regions in yellow boxes show the enhanced sharp features. This finding confirms that the channel attention mechanism can enhance the critical feature representations.

We also conduct an ablation study to quantitatively evaluate the factors benefiting the reconstruction of cloud-occluded regions, which are explained in Section 3. In this study, we consider the baseline results as the results from our proposed HS^{2}P without RBCA, which means that we use a scaling layer to replace the attention module in RBCA and bypass the hierarchical outputs to directly obtain the final output P_N. The attention module in the data fusion module is also discarded. Next, RBCA is incorporated in the second model. The third model develops on the second model with the hierarchical spectral outputs, which is trained with the spectral preserving loss. The last model is our complete model HS^{2}P with all improvements, which has hierarchical outputs and is trained with the custom collaborative optimization loss. The quantitative results of this study are presented

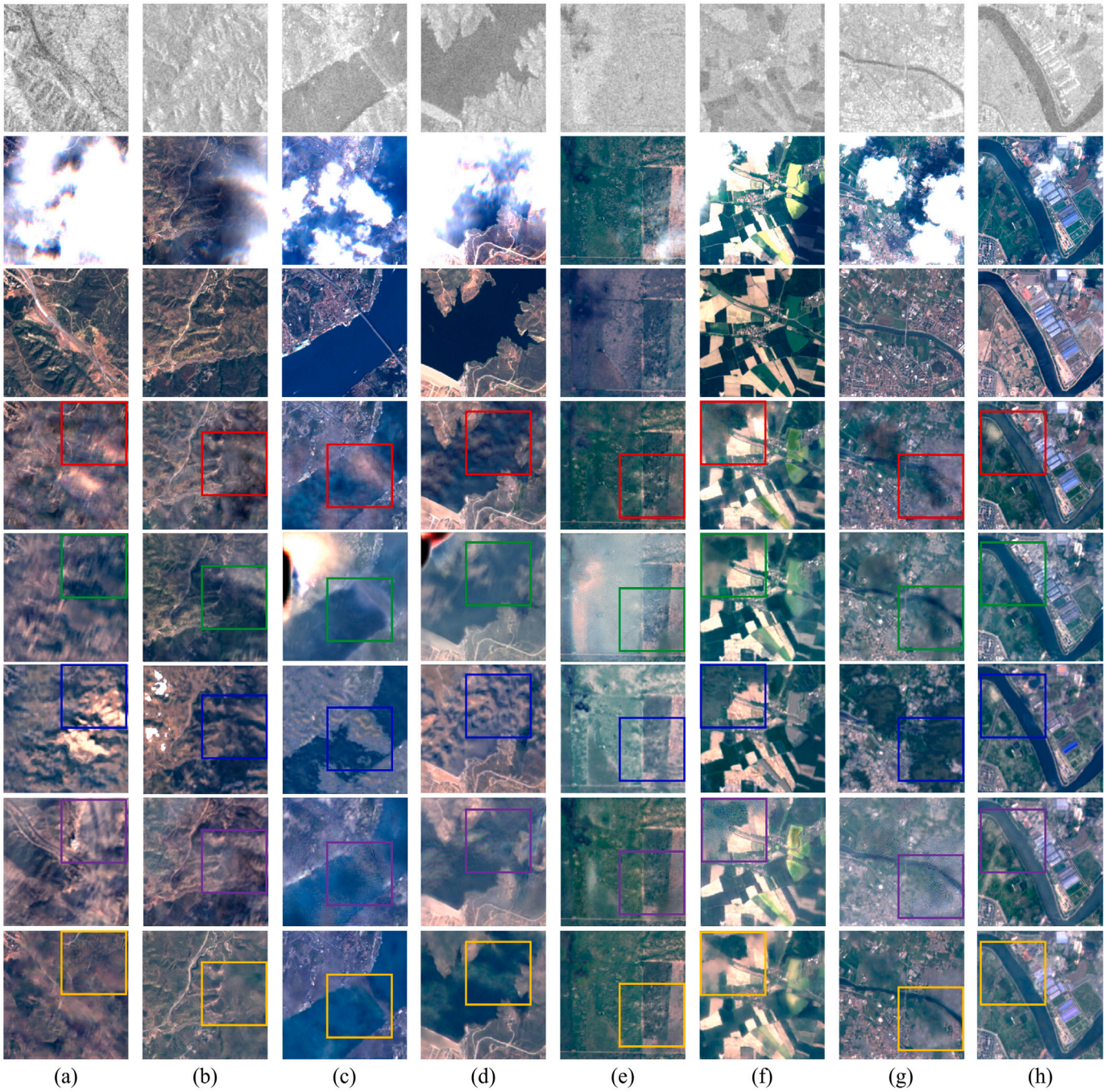

Fig. 5. Cloud and shadow removal results classified by land cover. The first row shows the Sentinel-1 SAR images, the second row shows the cloudy Sentinel-2 optical images, the third row shows the cloud-free Sentinel-2 optical images, the fourth row shows the results of the DSen2-CR model, the fifth row shows the results of the GAN-CR model, the sixth row shows the results of the AMGAN-CR model, the seventh row shows the results of the GLF-CR model and the last row shows the results of the HS^{2}P model.

in Table 2, where HSP and HST denote the hierarchical spectral outputs and hierarchical structural outputs, respectively. As shown in Table 2, the proposed HS^{2}P significantly improves the baseline results on MAE/RMSE/PSNR/SAM/SSIM by ∼11/15/5/8/1.4%, respectively, demonstrating its effectiveness in improving the reconstruction.

4.5. Analysis of the critical parameters

In the following section, the sensitivity analysis of experimental critical parameters is given. An analysis of ResGroups is provided in Section 4.5.1 to confirm the influence of ResGroups number N. Section 4.5.2 analyzes the effect of weight α between the spectral preserving loss and structural preserving loss on the model performance.

4.5.1. Analysis of the ResGroups number

We further investigate three models with different values of N to analyze the influence of the ResGroups number. The average quantitative results are presented in Fig. 8. As shown in Fig. 8, when $N = 4$ and $N = 6$, the average quantitative results are improved. However, the four ResGroups model and six ResGroups model show superiority in different metrics. For the choice of N, the four ResGroups model makes improvements on MAE/RMSE/PSNR by ∼1.4/1.0/0.9%, respectively, and the six ResGroups model makes improvements on

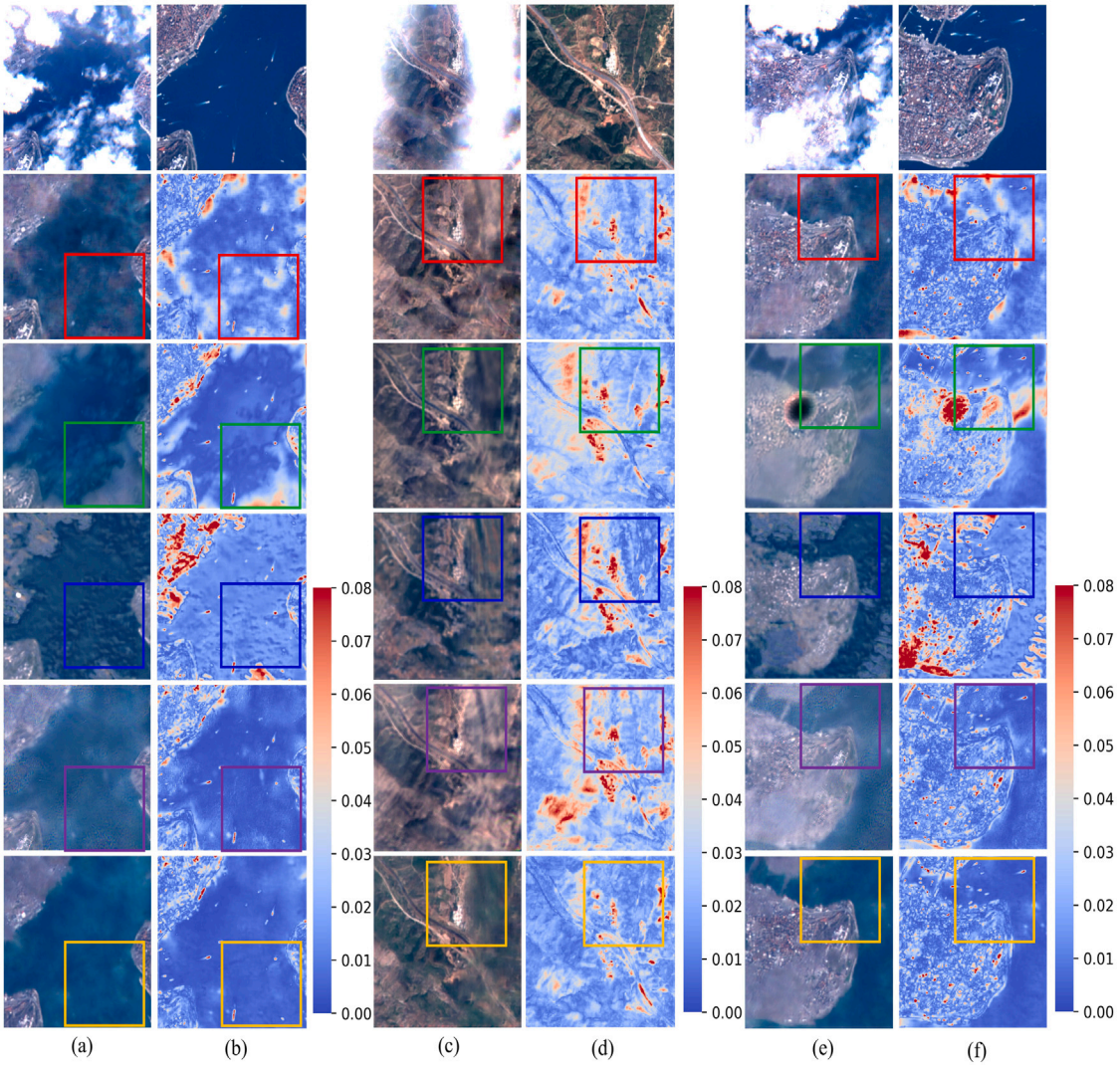

Fig. 6. Cloud and shadow removal results with absolute error maps. The first row shows the paired cloudy and cloud-free Sentinel-2 optical images, the second row shows the results of the DSen2-CR model, the third row shows the results of the GAN-CR model, the fourth row shows the results of the AMGAN-CR model, the fifth row shows the results of the GLF-CR model and the last row shows the results of the HS^{2}P model.

SAM/SSIM by ∼1.3/0.4% when comparing these two models, which shows the better performance of the four ResGroups model. Moreover, the six ResGroups model causes more time consumption in the training phase, and we suggest that it is unnecessary to sacrifice time for slight improvements in SAM and SSIM. Therefore, this study clarifies the selection of $N = 4$.

4.5.2. Analysis of the collaborative optimization loss weight

The regularization constant α is a weight to adjust the spectral preserving loss and structural preserving loss. It is also an important factor affecting the precision of the experimental results. To investigate the effect of α, we conduct a further study of α. There are three models to compare, which have α values of 0.25, 0.50 and 0.75. Fig. 9 shows the average quantitative results; the model with an α of 0.50 performs best on all the metrics.

Furthermore, we select some representative scenes with distinct boundaries of ground objects and display the cloud and shadow removal results that are predicted by the three models introduced above in Fig. 10. With an increase in α, the cloud and shadow removal results are more clearly and completely contoured but have worse preservation of spectral information, as the predicted images are more indistinctly colored compared with the ground truth. This phenomenon shows that α is proportional to the structural quality of the predicted images but inversely proportional to the spectral quality. With the above considerations, we use an α of 0.50 to simultaneously enhance the spectral and structural features.

4.6. Application on large-scale scenes

With the purpose of evaluating the robustness of the proposed method, we further resample some large-scale scenes that are not included in the used dataset and utilize our fully trained model to generate large-scale cloud removal results. The triples of large-scale images containing cloudy Sentinel-2 optical images, cloud-free Sentinel-2 optical images and Sentinel-1 SAR images are sampled with the size of 10 240 × 10 240 pixels. Additionally, we select scenes with various cloud types as well as land cover to guarantee the robust transfer capability of our model. As Fig. 11 shows, the scene in the first row contains waters and urban areas with thin cloud cover, the scene in the second row is mountains and the scene in the last row is mainly croplands which are both covered by opaquely thick clouds. Due to the limitation of memory capacity, the large-scale images are first clipped into the size of 1024 × 1024 pixels with strides of 128 pixels for prediction and the reconstructed patches are merged to produce the complete large cloud-free images, where we take the mean for the overlapping regions of the patches. It can be observed from Fig. 11 that our method

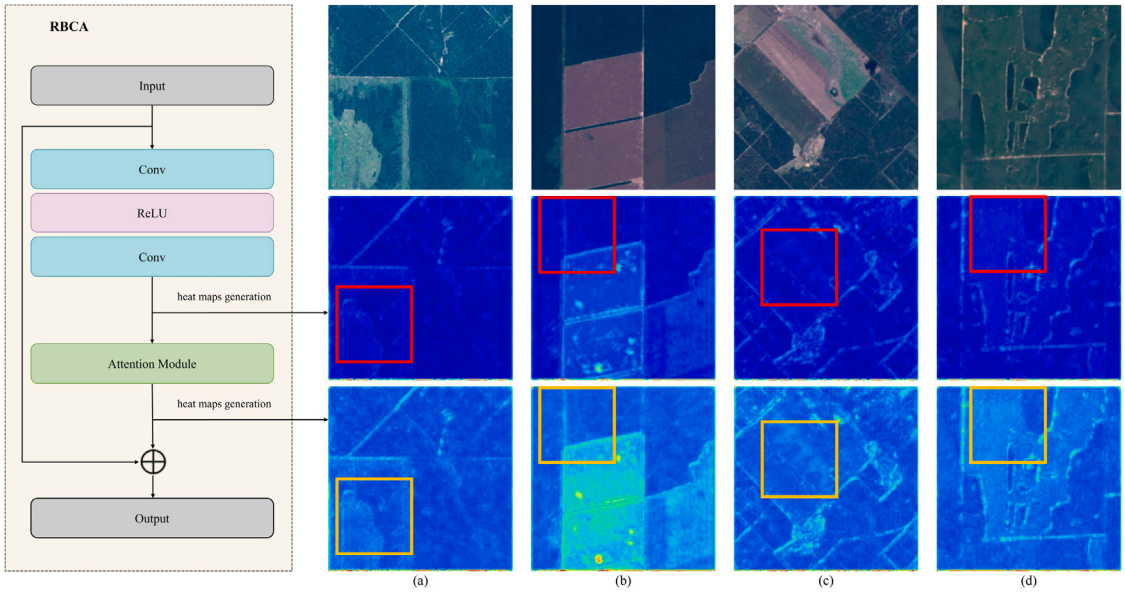

Fig. 7. Feature maps obtained before and after the attention module. The first row shows the cloud-free Sentinel-2 optical images, the second row shows the feature maps obtained before the attention module, and the third row shows the feature maps obtained after the attention module.

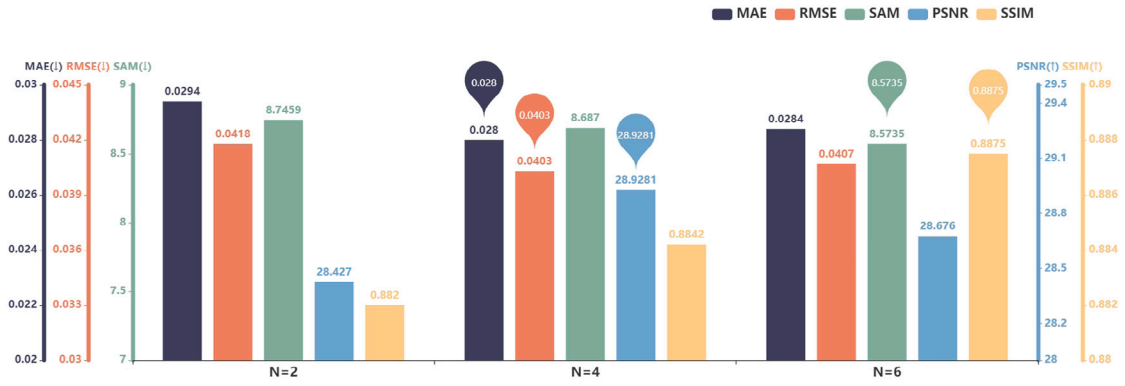

Fig. 8. Average quantitative results of models with different ResGroups number N when $\alpha = 0.25$.

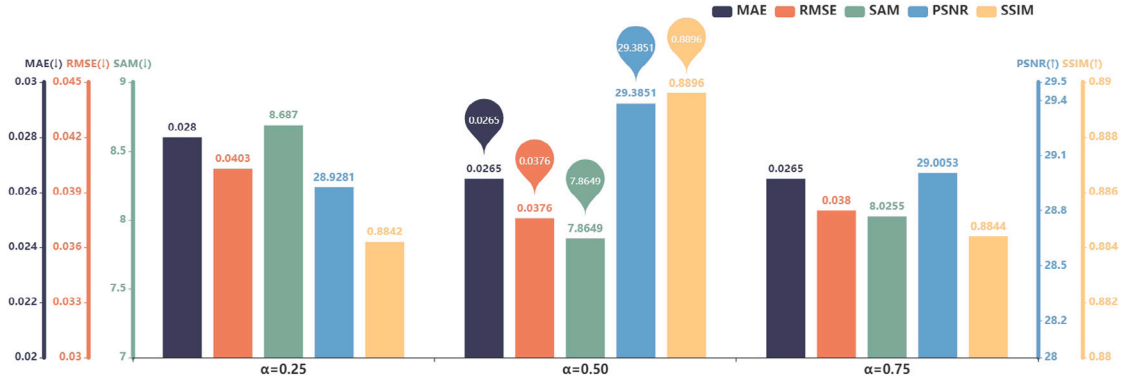

Fig. 9. Average quantitative results of models with different regularization constants α when $N = 4$.

generates cloud removal results of the resampled large-scale scenes in good quality, which proves its stability and robustness.

5. Conclusion

Optical remote sensing images are utilized in various applications. Nevertheless, optical images are often contaminated by clouds. As a basic step of processing images in the optical domain, cloud and shadow removal provides data support for continuous ground monitoring. Although SAR data can offer complementary contextual and structural information for cloud and shadow removal in optical remote sensing imagery, they are corrupted by speckle noise, which makes SAR-optical fusion ill-posed to generate cloud removal results with high quality. In this article, the cloud and shadow removal method HS^{2}P based on the fusion of Sentinel-2 optical data and Sentinel-1 SAR data is proposed.

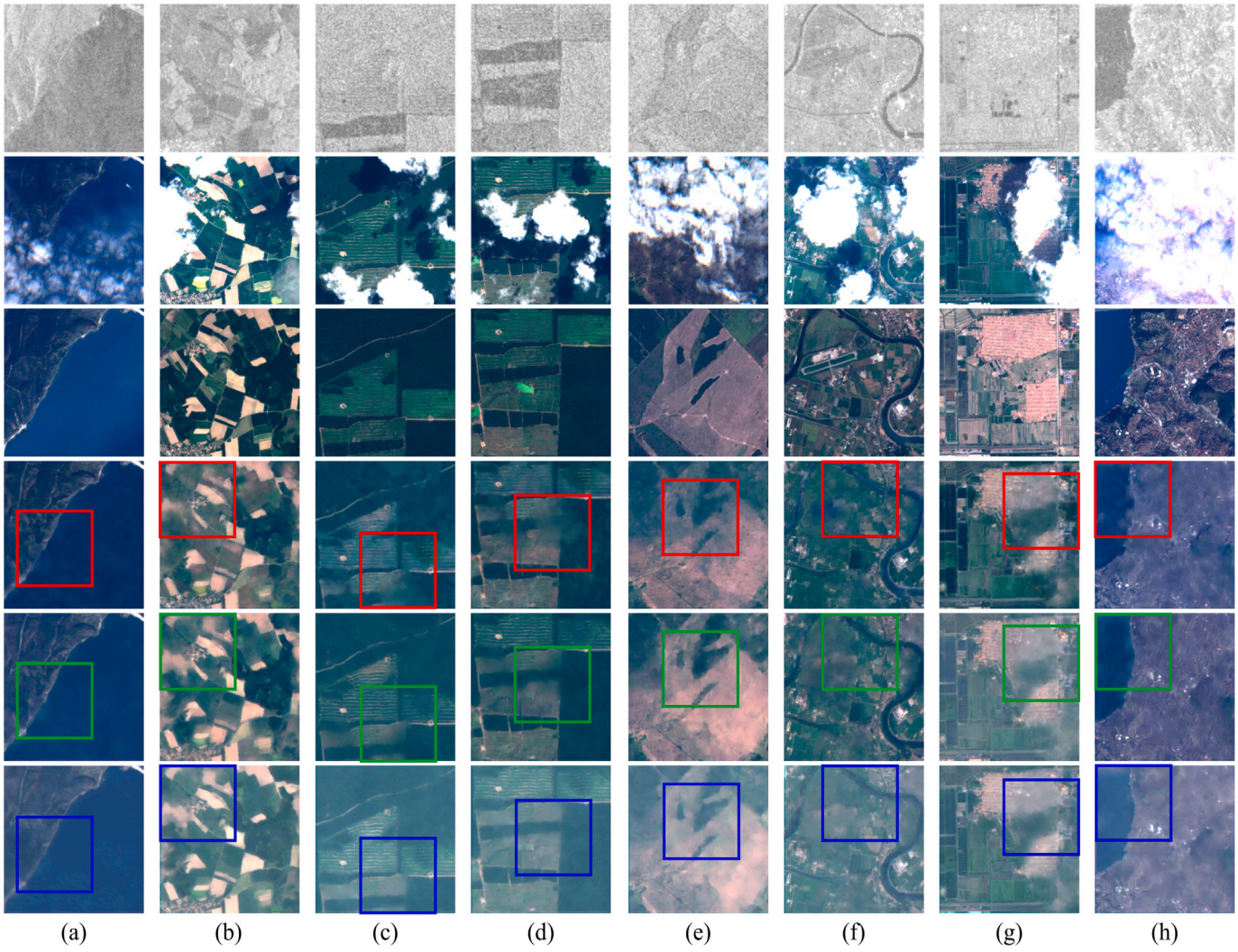

Fig. 10. Cloud and shadow removal results of models with different α. The first row shows the Sentinel-1 SAR images, the second row shows the cloudy Sentinel-2 optical images, the third row shows the cloud-free Sentinel-2 optical images, and Rows 4, 5 and 6 show the results of models with α values of 0.25, 0.50 and 0.75, respectively.

To progressively constrain the reconstruction at multiple levels of the network, we propose a deep hierarchical architecture. Furthermore, RBCA in HS^2P are embedded with a channel attention mechanism to pursue the adaptive selection of more informative features for fusion instead of equally treating channelwise features of multimodal imagery. We train the proposed model with the custom collaborative optimization loss to make the network generate cloud removal results with not only a fine appearance but also explicit outlines. Then, two state-of-the-art cloud and shadow removal methods are compared with our method on the SEN12MS-CR dataset. The experimental results demonstrate that our method, which can handle various cloud types and reconstruct diverse land covers, achieves significant improvements. We also conduct an ablation study and prove the effectiveness of our contributions. For future work, we will try more advanced deep network architectures to improve the performance on cloud and shadow removal and will consider the fusion of more remote sensing data to enhance the reconstruction of cloudy regions in a further step. Utilizing the polarization information of SAR data and the properties of SAR imaging system will be taken into consideration as well to make better use of SAR data as complementary data source.

CRediT authorship contribution statement

Yansheng Li: Conceptualization of this study, Methodology, Writing. **Fanyi Wei:** Methodology, Experiment, Writing. **Yongjun Zhang:** Conceptualization of this study, Revised the manuscript. **Wei Chen:** Experiment, Revised the manuscript. **Jiayi Ma:** Conceptualization of this study, Revised the manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request

Acknowledgments

This work was supported in part by the State Key Program of the National Natural Science Foundation of China under Grant 42030102, the National Natural Science Foundation of China under Grant 41971284, the National Natural Science Foundation of China under Grant 62276192, the Fundamental Research Funds for the Central Universities, China under grant 2042022kf1201, the Zhizhuo Research Fund on Spatial-Temporal Artificial Intelligence, China under grant ZZJJ202210, and the Wuhan University-Huawei Geoinformatics Innovation Laboratory, China.

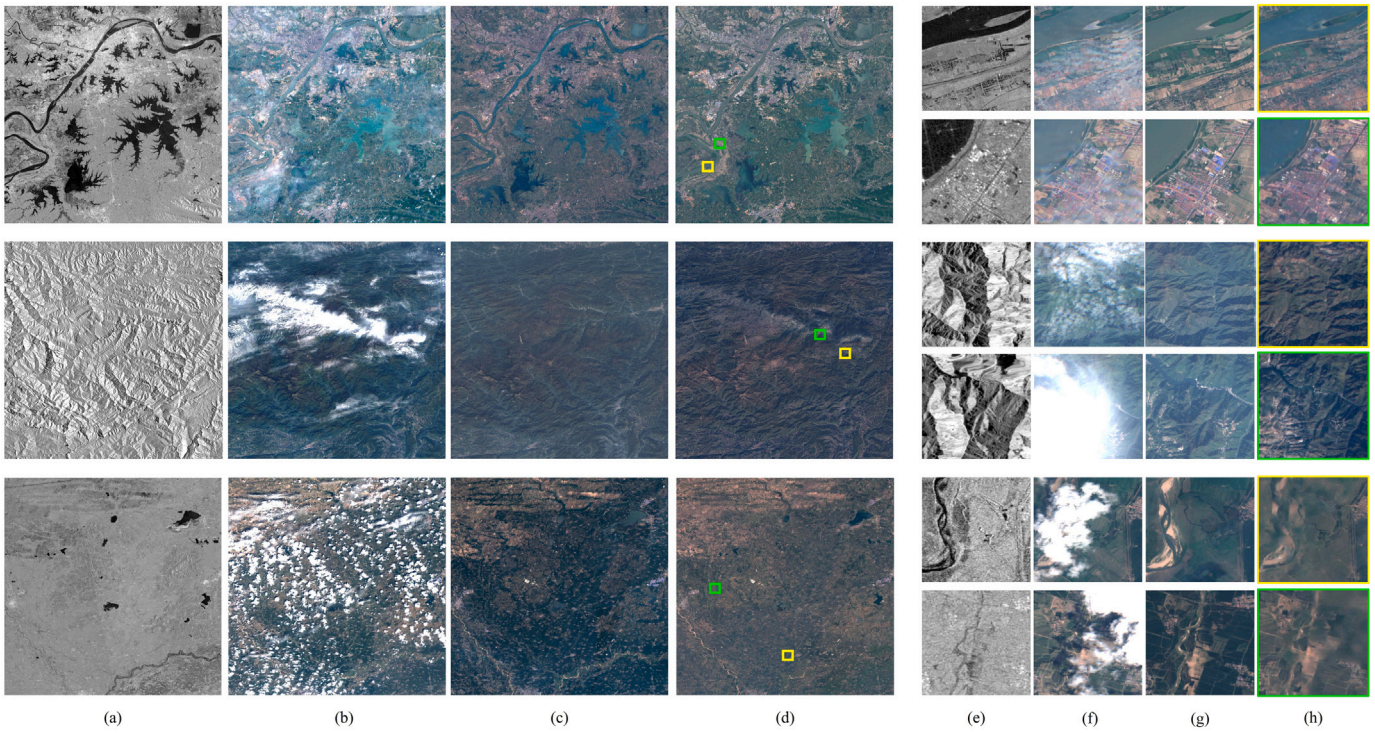

Fig. 11. Cloud and shadow removal results of large-scale scenes. Column *a* shows the Sentinel-1 SAR images. Column *b* shows the cloudy Sentinel-2 optical images. Column *c* shows the cloud-free Sentinel-2 optical images. Column *d* shows the results of the HS^{2}P model. Columns *e* – *h* show the details in Yellow and Green boxes of the corresponding images in the left half.

References

[1] M. Wieland, Y. Li, S. Martinis, Multi-sensor cloud and cloud shadow segmentation with a convolutional neural network, Remote Sens. Environ. 230 (2019) 111203.

[2] Y. Zhang, W.B. Rossow, A.A. Lacis, V. Oinas, M.I. Mishchenko, Calculation of radiative fluxes from the surface to top of atmosphere based on ISCCP and other global data sets: Refinements of the radiative transfer model and the input data, J. Geophys. Res.: Atmos. 109 (D19) (2004).

[3] J. Ju, D.P. Roy, The availability of cloud-free Landsat ETM+ data over the conterminous United States and globally, Remote Sens. Environ. 112 (3) (2008) 1196–1211.

[4] Q. Cheng, H. Shen, L. Zhang, P. Li, Inpainting for remotely sensed images with a multichannel nonlocal total variation model, IEEE Trans. Geosci. Remote Sens. 52 (1) (2013) 175–187.

[5] I. Gladkova, M.D. Grossberg, F. Shahriar, G. Bonev, P. Romanov, Quantitative restoration for MODIS band 6 on Aqua, IEEE Trans. Geosci. Remote Sens. 50 (6) (2011) 2409–2416.

[6] F. Meng, X. Yang, C. Zhou, Z. Li, A sparse dictionary learning-based adaptive patch inpainting method for thick clouds removal from high-spatial resolution remote sensing imagery, Sensors 17 (9) (2017) 2130.

[7] M. Xu, X. Jia, M. Pickering, S. Jia, Thin cloud removal from optical remote sensing images using the noise-adjusted principal components transform, ISPRS J. Photogramm. Remote Sens. 149 (2019) 215–225.

[8] Y. Chen, W. He, N. Yokoya, T.-Z. Huang, Blind cloud and cloud shadow removal of multitemporal images based on total variation regularized low-rank sparsity decomposition, ISPRS J. Photogramm. Remote Sens. 157 (2019) 93–107.

[9] X. Li, H. Shen, L. Zhang, H. Zhang, Q. Yuan, G. Yang, Recovering quantitative remote sensing products contaminated by thick clouds and shadows using multitemporal dictionary learning, IEEE Trans. Geosci. Remote Sens. 52 (11) (2014) 7086–7098.

[10] J. Lin, T.-Z. Huang, X.-L. Zhao, M. Ding, Y. Chen, T.-X. Jiang, A blind cloud/shadow removal strategy for multi-temporal remote sensing images, in: 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, IEEE, 2021, pp. 4656–4659.

[11] H. Shen, J. Wu, Q. Cheng, M. Aihemaiti, C. Zhang, Z. Li, A spatiotemporal fusion based cloud removal method for remote sensing images with land cover changes, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 12 (3) (2019) 862–874.

[12] M. Schmitt, L. Hughes, C. Qiu, X.X. Zhu, Aggregating cloud-free Sentinel-2 images with Google earth engine, PIA19: Photogramm. Image Anal. (2019) 145–152.

[13] C. Grohnfeldt, M. Schmitt, X. Zhu, A conditional generative adversarial network to fuse SAR and multispectral optical data for cloud removal from Sentinel-2 images, in: IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium, IEEE, 2018, pp. 1726–1729.

[14] J. Gao, H. Zhang, Q. Yuan, Cloud removal with fusion of SAR and optical images by deep learning, in: 2019 10th International Workshop on the Analysis of Multitemporal Remote Sensing Images (MultiTemp), IEEE, 2019, pp. 1–3.

[15] H. Zhang, H. Xu, X. Tian, J. Jiang, J. Ma, Image fusion meets deep learning: A survey and perspective, Inf. Fusion 76 (2021) 323–336.

[16] J. Ma, L. Tang, F. Fan, J. Huang, X. Mei, Y. Ma, SwinFusion: Cross-domain long-range learning for general image fusion via swin transformer, IEEE/CAA J. Autom. Sin. 9 (7) (2022) 1200–1217.

[17] L. Tang, Y. Deng, Y. Ma, J. Huang, J. Ma, SuperFusion: A versatile image registration and fusion network with semantic awareness, IEEE/CAA J. Autom. Sin. 9 (12) (2022) 2121–2137.

[18] Z. Wang, Y. Ma, Y. Zhang, Review of pixel-level remote sensing image fusion based on deep learning, Inf. Fusion (2022).

[19] S.C. Kulkarni, P.P. Rege, Pixel level fusion techniques for SAR and optical images: A review, Inf. Fusion 59 (2020) 13–29.

[20] F.N. Darbaghshahi, M.R. Mohammadi, M. Soryani, Cloud removal in remote sensing images using generative adversarial networks and SAR-to-optical image translation, IEEE Trans. Geosci. Remote Sens. 60 (2021) 1–9.

[21] A. Meraner, P. Ebel, X.X. Zhu, M. Schmitt, Cloud removal in Sentinel-2 imagery using a deep residual neural network and SAR-optical data fusion, ISPRS J. Photogramm. Remote Sens. 166 (2020) 333–346.

[22] W. Li, Y. Li, D. Chen, J.C.-W. Chan, Thin cloud removal with residual symmetrical concatenation network, ISPRS J. Photogramm. Remote Sens. 153 (2019) 137–150.

[23] Q. Yang, G. Wang, Y. Zhao, X. Zhang, G. Dong, P. Ren, Multi-scale deep residual learning for cloud removal, in: IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium, IEEE, 2020, pp. 4967–4970.

[24] J. Bermudez, P. Happ, D. Oliveira, R. Feitosa, Sar to optical image synthesis for cloud removal with generative adversarial networks, ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci. 4 (1) (2018).

[25] P. Isola, J.-Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1125–1134.

[26] J. Gao, Q. Yuan, J. Li, H. Zhang, X. Su, Cloud removal with fusion of high resolution optical and SAR images using generative adversarial networks, Remote Sens. 12 (1) (2020) 191.

[27] V. Sarukkai, A. Jain, B. Uzkent, S. Ermon, Cloud removal from satellite images using spatiotemporal generator networks, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020, pp. 1796–1805.

[28] P. Ebel, M. Schmitt, X.X. Zhu, Cloud removal in unpaired Sentinel-2 imagery using cycle-consistent GAN and SAR-optical data fusion, in: IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium, IEEE, 2020, pp. 2065–2068.
[29] Y. Gan, T. Xiang, H. Liu, M. Ye, Learning-aware feature denoising discriminator, Inf. Fusion 89 (2023) 143–154.
[30] D.M. Vo, D.M. Nguyen, T.P. Le, S.-W. Lee, HI-GAN: A hierarchical generative adversarial network for blind denoising of real photographs, Inform. Sci. 570 (2021) 225–240.
[31] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[32] K. He, X. Zhang, S. Ren, J. Sun, Identity mappings in deep residual networks, in: European Conference on Computer Vision, Springer, 2016, pp. 630–645.
[33] X. Liu, Q. Liu, Y. Wang, Remote sensing image fusion based on two-stream fusion network, Inf. Fusion 55 (2020) 1–15.
[34] Q. He, X. Sun, Z. Yan, K. Fu, DABNet: Deformable contextual and boundary-weighted network for cloud detection in remote sensing images, IEEE Trans. Geosci. Remote Sens. 60 (2021) 1–16.
[35] M. Xu, F. Deng, S. Jia, X. Jia, A.J. Plaza, Attention mechanism-based generative adversarial networks for cloud removal in Landsat images, Remote Sens. Environ. 271 (2022) 112902.
[36] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.
[37] C. Duanmu, J. Zhu, The image super-resolution algorithm based on the dense space attention network, IEEE Access 8 (2020) 140599–140606.
[38] S. Woo, J. Park, J.-Y. Lee, I.S. Kweon, Cbam: Convolutional block attention module, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 3–19.
[39] Y. Zhou, W. Jing, J. Wang, G. Chen, R. Scherer, R. Damaševičius, MSAR-DefogNet: Lightweight cloud removal network for high resolution remote sensing images based on multi scale convolution, IET Image Process. 16 (3) (2022) 659–668.
[40] X. Wen, Z. Pan, Y. Hu, J. Liu, An effective network integrating residual learning and channel attention mechanism for thin cloud removal, IEEE Geosci. Remote Sens. Lett. 19 (2022) 1–5.
[41] Q. He, X. Sun, Z. Yan, B. Li, K. Fu, Multi-object tracking in satellite videos with graph-based multitask modeling, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–13.
[42] Q. He, X. Sun, W. Diao, Z. Yan, D. Yin, K. Fu, Transformer-induced graph reasoning for multimodal semantic segmentation in remote sensing, ISPRS J. Photogramm. Remote Sens. 193 (2022) 90–103.
[43] F. Xu, Y. Shi, P. Ebel, L. Yu, G.-S. Xia, W. Yang, X.X. Zhu, GLF-CR: SAR-enhanced cloud removal with global–local fusion, ISPRS J. Photogramm. Remote Sens. 192 (2022) 268–278.
[44] P. Dai, S. Ji, Y. Zhang, Gated convolutional networks for cloud removal from bi-temporal remote sensing images, Remote Sens. 12 (20) (2020) 3427.
[45] A. Kuznetsov, M. Gashnikov, Remote sensing image inpainting with generative adversarial networks, in: 2020 8th International Symposium on Digital Forensics and Security, ISDFS, IEEE, 2020, pp. 1–6.
[46] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, Y. Fu, Image super-resolution using very deep residual channel attention networks, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 286–301.
[47] C. Ma, Y. Rao, Y. Cheng, C. Chen, J. Lu, J. Zhou, Structure-preserving super resolution with gradient guidance, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 7769–7778.
[48] Z. Zhang, K. Gao, J. Wang, L. Min, S. Ji, C. Ni, D. Chen, Gradient enhanced dual regression network: Perception-preserving super-resolution for multi-sensor remote sensing imagery, IEEE Geosci. Remote Sens. Lett. 19 (2021) 1–5.
[49] A. Natekin, A. Knoll, Gradient boosting machines, a tutorial, Front. Neurorobot. 7 (2013) 21.
[50] P. Ebel, A. Meraner, M. Schmitt, X.X. Zhu, Multisensor data fusion for cloud removal in global and all-season sentinel-2 imagery, IEEE Trans. Geosci. Remote Sens. 59 (7) (2020) 5866–5878.
[51] F.A. Kruse, A. Lefkoff, J. Boardman, K. Heidebrecht, A. Shapiro, P. Barloon, A. Goetz, The spectral image processing system (SIPS)—interactive visualization and analysis of imaging spectrometer data, Remote Sens. Environ. 44 (2–3) (1993) 145–163.
[52] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.
[53] C. Lanaras, J. Bioucas-Dias, S. Galliani, E. Baltsavias, K. Schindler, Super-resolution of Sentinel-2 images: Learning a globally applicable deep neural network, ISPRS J. Photogramm. Remote Sens. 146 (2018) 305–319.