MMCANet A Multimodal and Cross-Attention Network for Cloud Removal and Exploration of Progressive Remote Sensing Images Restoration Algorithm

Yejian Zhou, *Member, IEEE*, Jiahui Suo, Yachen Wang, Jie Su, Wen Xiao, *Member, IEEE*, Zhen Hong, *Member, IEEE*, Rajiv Ranjan, *Fellow, IEEE*, Lizhe Wang, *Fellow, IEEE*, and Zhenyu Wen, *Senior Member, IEEE*

***Abstract*— In Earth observation, cloud severely affects the interpretation of optical satellites generated high-resolution images. Cloud-free optical images are vital for downstream tasks such as semantic segmentation and object detection. Thus, the elimination of clouds from optical imagery has emerged as a significant topic in remote sensing. Currently, most existing methods are proposed to leverage the texture information from auxiliary synthetic aperture radar (SAR) images to restore cloud-free images via direct channel merging. However, such a unified feature extraction approach often neglects the inherent distribution disparity between SAR and optical images—the result of differing imaging principles-potentially leading to significant feature loss. To this end, we introduce a network by jointing SAR and optical images multimodal and cross-attention network (MMCANet) to effectively extract multiscale contextual features from SAR imagery and integrate them with optical features. Specifically, instead of simple concatenation of the channels of SAR and optical images, we obtain high-dimensional features from them through independent feature extractors. The integration of these features is facilitated by a cross-attention mechanism that provides a more fine-grained amalgamation of information. Meanwhile, an atrous spatial pyramid pooling (ASPP) module is introduced into the integration of high-level features, which captures multiscale contextual information around clouded areas. In addition, we propose four advanced remote sensing image restoration algorithms that approach image restoration as a series of subtasks, gradually eliminating clouds to enhance performance. Comprehensive assessments show that MMCANet performs well on the SEN 12 MS-CR dataset with peak signal-to-noise ratio (PSNR) of 39.8871, structural similarity index (SSIM) of 0.9672, mean absolute error (MAE) of 0.0081, and spectral angle mapper (SAM) of 2.9884.**

***Index Terms*— Cloud removal, deep learning, feature fusion, image restoration.**

Received 27 August 2024; revised 18 November 2024 and 15 February 2025; accepted 26 March 2025. Date of publication 1 April 2025; date of current version 14 April 2025. This work was supported in part by the National Natural Sciences Foundation of China under Grant 62471438 and Grant 62472387, in part by Zhejiang Provincial Natural Science Foundation of China under Grant LY23F010012, in part by China Postdoctoral Science Foundation under Grant 2023M743403, and in part by Zhejiang Provincial Natural Science Foundation of Major Program (Youth Original Project) under Grant LDQ24F020001. *(Corresponding authors: Jie Su; Zhenyu Wen.)*

Yejian Zhou and Yachen Wang are with the College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China.

Jiahui Suo, Jie Su, and Zhen Hong are with the Institute of Cyberspace Security, Zhejiang University of Technology, Hangzhou 310023, China (e-mail: jieamsu@gmail.com).

Wen Xiao and Lizhe Wang are with the School of Geography and Information Engineering, China University of Geosciences, Wuhan 430074, China.

Rajiv Ranjan is with the School of Computing, Newcastle University, NE4 5TG Newcastle Upon Tyne, U.K.

Zhenyu Wen is with the Institute of Cyberspace Security and the College of Information Engineering, Zhejiang University of Technology, Hangzhou, Zhejiang 310023, China, and also with the Department of Electronics Information Science and Technology, University of Science and Technology of China, Hefei, Anhui 230025, China (e-mail: zhenyuwen@zjut.edu.cn).

Digital Object Identifier 10.1109/TGRS.2025.3556560

I. Introduction

As REMOTE sensing technology advances, high-resolution optical images from satellites are increasingly used to support various Earth observation applications, including micro-object detection, surveying, and disaster monitoring [1], [2], [3]. However, clouds present a significant challenge in interpreting these spaceborne optical images. According to existing research, over 55% of land areas in these images are obscured by clouds [4]. The concealment of image contents can be caused by thick clouds, while even thin translucent clouds can significantly distort the ground below, thereby greatly affecting the usefulness of satellite images [5]. Therefore, reconstructing high-quality images from cloud-contaminated degraded images is an indispensable preprocessing step for applications. Removing thin clouds restores partially obscured targets, making their outlines clearer and reducing the likelihood of misjudgment in subsequent tasks. Removing thick clouds supports specific sequential observation tasks in the same area.

Due to the loss of texture information in areas obscured by clouds, the task of cloud removal is markedly ill-posed. Traditional cloud-removal approaches typically utilize the information from cloud-free regions on current images in spatial-based methods [6] or past-time images in temporal-based methods [7] to estimate the areas under cloud cover. However, when relying on the current cloud-free areas to restore occluded pixels, the approach might fail due to extensive cloud coverage. On the other hand, using past-time cloud-free areas can be affected significantly by landscape changes, thereby greatly impacting the final restored images.

Benefit from the synthetic aperture radar (SAR) images, which are unaffected by cloud coverage due to their superior penetrability and ability to measure backscatter, significantly mitigate these challenges. By leveraging the

complementary information between the auxiliary images and the corresponding optical images, these methods [8], [9] can reconstruct the contaminated image. However, due to their distinct imaging mechanisms, SAR and optical images manifest different characteristics of the observed objects, thereby creating a substantial domain gap between them. Existing methods frequently depend on the simplistic superimposition of SAR and optical image channels for data fusion. This approach could result in inadequate information complementarity and consequently lead to unstable model performance [10], [11], [12].

Recently, [13] proposed a method to integrate SAR and optical modalities through a dual-stream global–local fusion, which significantly minimizes the domain gap effect arising from direct channel superimposition, yielding substantial performance improvements. However, the image restoration process employed by these methods is typically executed in a single-stage fashion and often depends on intricate network architectures. The single-stage network's nonlinear representation capacity is frequently insufficient, making it challenging to reconstruct the rich texture information present in images under complex scenarios.

To solve the aforementioned problems and limitations, we propose a method by jointing SAR and optical images multimodal and cross-attention network (MMCANet) to recover the missing regions in optical images. Specifically, we leverage a dual-encoder structure to extract contextual information from both SAR and optical images and employ a cross-attention mechanism [14] to integrate features at low and high-level representations, thus enabling multimodal data interaction. This module seeks to maximize the compensatory capability of SAR images for texture details. Furthermore, in light of the stochastic nature of cloud coverage, we incorporate an atrous spatial pyramid pooling (ASPP) [15] on the fused high-level features to extract multiscale contextual information near the cloud regions. This facilitates improved restoration of optical images across a range of cloud coverage levels. To further optimize the performance of MMCANet, we develop four progressive restoration architectures and investigate their potential for cloud removal. The restoration process is decomposed into several sub-tasks, gradually restoring high-resolution images. As such, the proposed algorithm is capable of better integrating multimodal data, resulting in high-quality, cloud-free images.

The contributions of our work can be summarized as follows.

1) We propose a network for jointing SAR and optical images, MMCANet. It extensively explores the beneficial role of SAR in restoring reliable texture details and maintaining the global consistency of the reconstructed images, thus facilitating effective reconstruction of areas obscured by cloud cover.
2) We combine the multimodal cross-attention module and the ASPP module in our dual-stream network to enhance the transmission of complementary information embedded in SAR images and the global interaction between contexts. This ensures that the structure of the recovered regions remains consistent with the remaining cloud-free areas, while also generating reliable texture details.
3) We have developed a multistage remote sensing image restoration framework and its variants, specialized for high-quality cloud removal. We experimentally validated their effectiveness and conducted a detailed analysis, thereby providing a research direction for future studies.

Organization. We propose the details of MMCANet and four multistage image restoration algorithms in Section III. All experimental results are given in Section IV. In Section V, we analyze the comparison results with other baselines in detail. Finally, we summarize this article in Section VI.

II. Related Work

A. Single-Stage Methods for Cloud Removal

The single-stage architecture is commonly employed in cloud removal networks, employing diverse functional components to enhance performance. Enomoto et al. [16] harnessed the substantial generative capabilities of conditional generative adversarial networks (cGAN) for thin cloud removal. Leveraging both NIR and optical images as inputs, their approach effectively eliminates clouds from visible light images when they are perceptible in the NIR images. Pan [17] introduced the spatial attention mechanism into the task of cloud removal and proposed a spatial attention generative adversarial network (SpA GAN) which only uses SAR images as input. This method focuses the network's attention more on cloud regions, thereby greatly improving the quality of the restored image. Bermudez et al. [10] employed cGANs to establish the relationship between optical and SAR images, using SAR to obtain optical images directly. However, the inherent differences between the two modalities, the generated optical images lack quality assurance. Grohnfeldt et al. [11] proposed a novel method based on cGAN for cloud removal by fusing images of the two modalities. They enhanced the ability to extract features and fuse multimodal data. Gao et al. [18] performed cloud removal by two-step. Initially, they generated simulated optical images using SAR. Subsequently, they combined simulated optical, optical, and SAR images to restore regions covered by clouds using GAN. Meraner et al. [12] employed a deep residual neural network to remove clouds by concatenating SAR and optical images. This fusion method may result in the loss of local information and is ineffective in reconstructing regions with significant cloud coverage.

In order to effectively leverage the complementary information inherent in SAR and optical images, [13] devised a dual-stream cloud removal (GLF-CR) algorithm aimed at extracting both global and local features from SAR and optical datasets while facilitating feature fusion. Integration of the Swin transformer and attention mechanism significantly bolsters the model's feature extraction capabilities. Nonetheless, it demonstrates a tendency toward overcorrection in tone prediction, leading to an overall darkened appearance in the resulting images. Han et al. [19] introduced a novel transformer-based network for feature extraction from the fusion of SAR and optical data. However, their approach

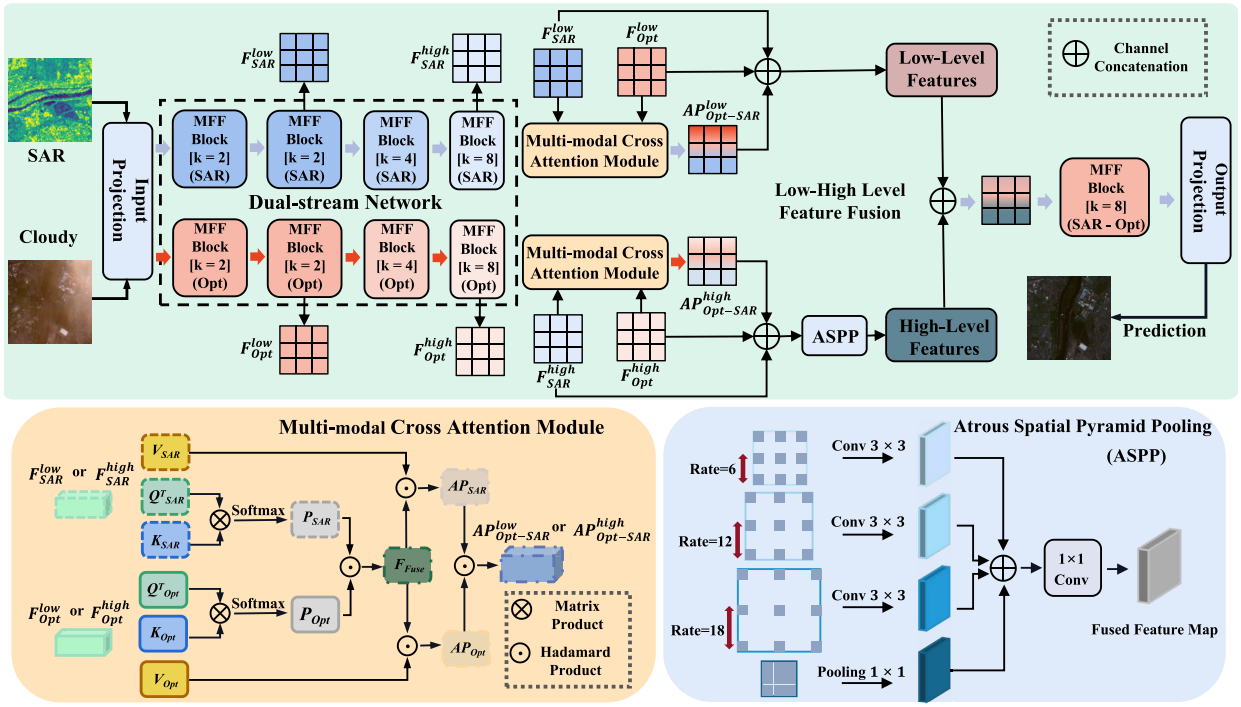


Fig. 1. Overall framework of our proposed model, MMCANet. The model consists of a dual-stream encoder and decoder, with the multiaxis feature fusion (MFF) blocks being crucial components. The encoder is illustrated in four encoding stages, each of which stacks a different number of MFF blocks. Similarly, the decoder also shows the number of stacked MFF blocks. k represents the MFF block numbers for each stage in the encoder and decoder. $F_{\text{SAR}}^{\text{high}}$ and $F_{\text{Opt}}^{\text{high}}$ represent the encoded high-level features while $F_{\text{SAR}}^{\text{low}}$ and $F_{\text{Opt}}^{\text{low}}$ represent the encoded low-level features. Key components of the model also include the multimodal cross attention module and ASPP.

merely concatenates SAR and optical channels without delving deeply into the feature fusion intricacies between SAR and optical imagery.

B. Multistage Methods for Image Restoration

Currently, the utilization of multistage structures in cloud removal research remains limited, despite its widespread application in computer vision. Prior studies [20], [21], [22] have demonstrated that multistage networks can outperform single-stage counterparts in sophisticated vision tasks such as pose estimation and action segmentation. Filtjens et al. [21] proposed a multistage spatial–temporal graph convolutional network (MS-GCN) that replaced the initial stage of temporal convolution with spatial graph convolution.

With advancements in deep learning, multistage models are no longer limited to high-level tasks but have been diversified and applied to various downstream tasks. Bai et al. [23], Yan et al. [24], and Manu [25] have achieved better performance by using multistage models to accomplish tasks such as dehazing, deblurring, denoising, and compound restoration. For example, [23] introduced MSPnet, a network composed of three denoising stages. Each stage comprises a parallel structure consisting of an encoder–decoder and a single-scale branch. The network decomposes the denoising into multiple sub-tasks, allowing for the gradual removal of noise through progressive steps.

III. Method

This section is composed of four parts: 1) single-stage image restoration; 2) multistage progressive image restoration (MPIR); 3) optimization objective of MMCANet; and 4) evaluation metrics.

A. Single-Stage Image Restoration

Fig. 1 illustrates the proposed MMCANet (single-stage image restoration variant). The proposed network employs a traditional encoder–decoder architecture to restore degraded images by harnessing multiscale feature extraction and fusion. Specifically, the cloudy and SAR images are first fed into the proposed dual-stream network to generate diverse levels of features for the subsequent fusion module. The low- and high-level features are then input into the cross-attention module to generate two joint attention maps $AP_{Opt-SAR}^{\text{low}}$ and $AP_{Opt-SAR}^{\text{high}}$. To integrate the aforementioned high-order and low-order features for image restoration, we introduce a low-high level feature fusion module. Subsequently, the optical-SAR fused feature is utilized as input to the decoder for image restoration. More details about these modules will be given in Sections III-A1–III-A3.

1) Dual-Stream Network: Previous research often leverages early fusion of SAR and optical images (i.e., concatenating the images prior to feature extraction) to perform modality integration. However, due to the domain gap between SAR and optical modalities, the integrated feature extractor cannot effectively extract mode-specific features, leading to unstable cloud removal outcomes. Thus, to amplify the use of complementary information inherent in SAR images, we utilize a dual-stream structured network to perform mode-specific feature extraction. Specifically, the dual-stream

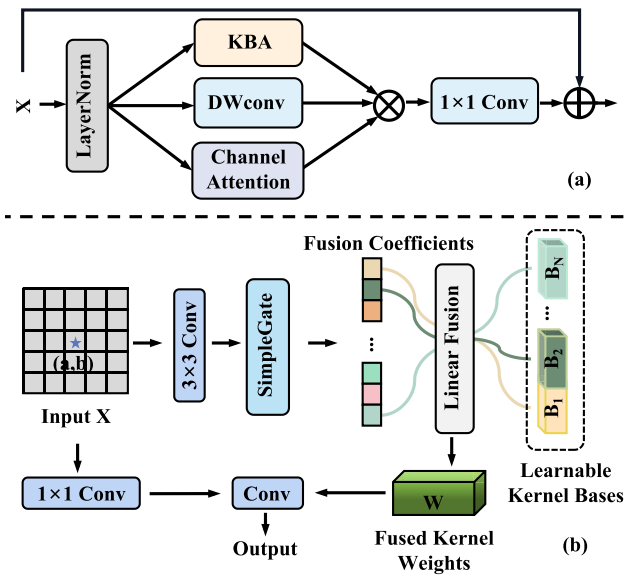


Fig. 2. Pipeline of MFF block. (a) Details of the MFF Block. It concludes three branches. (b) Kernel basis attention (KBA) module in MFF block.

network is composed of numerous multiaxis feature fusion (MFF) blocks [26] [as presented in Fig. 2 (a)], each having distinct kernel sizes for the optical and SAR modalities. The MFF block enables the network to extract detailed channel and spatial information through a diverse set of kernel bases, thereby facilitating the adaptive modeling of various local structures.

The MFF block initially encodes the input features through layer normalization, which aims to stabilize the training process and facilitate spatial information aggregation. Subsequently, three operators—namely, depth-wise convolution (DWconv) [27], channel attention [28], and Kernel basis attention (KBA) [26]—are utilized to extract comprehensive mode features. The 3×3 DWconv is designed to capture spatially invariant features, while channel attention serves to modulate feature channels. KBA module is intended to manage spatial features adaptively, which is illustrated in Fig. 2(b). Specifically, at each pixel position, the KBA module utilizes the learned kernel bases to compute localized feature representations and predicts a pixel-wise coefficient vector. This vector is subsequently used to conduct a linear combination of various kernel bases, thereby generating the final feature representation. The pixel-wise coefficient vector is learned during training and predicted for each pixel position, allowing adaptive aggregation of spatial information and improving model performance. At last, a residual shortcut with point-wise multiplication is applied, promoting training convergence and directly fusing diverse features from the three branches. To further enhance the MFF's nonlinear transformation capability, we attempt to replace the 3×3 convolution in DWconv with a 5×5 convolution. Although Table II show s stronger restoration capability, the computational workload significantly increased.

2) Multimodal Cross Attention Module: The SAR and optical images have different physical properties and imaging mechanisms result in different information expression formats. SAR can provide detailed structural information, including edge details and texture features of the contaminated area, while optical images can offer rich spectral information to achieve higher spectral fidelity in the generated images. Thus, leveraging both SAR and optical features can provide intensive fine-grained information for image restoration task and bring high-quality results. Inspired by that, we introduce a multimodal cross-attention module to perform the cross-modal information integration based on the extracted domain-specific features. The proposed multimodal cross-attention module aims to take the extracted low-level and high-level features from the dual-stream network as the input to generate high-dimensional joint features via the cross-attention mechanism (as illustrated in Fig. 1). Specifically, the single modality attention mechanism initially transforms the single modality features, F_{Opt} and F_{SAR}, into three distinct feature maps: Q_{Opt}, Q_{SAR}, K_{Opt}, K_{SAR}, and V_{Opt}, V_{SAR}, using 1×1 convolutions. Subsequently, the transpose of the query feature Q is multiplied by the key feature K. This product is then passed through a softmax layer to derive the self-attention map P. The equations to extract the low-level self-attention maps for both optical and SAR modalities are as follows:

$$P_{\text{Opt}}^{\text{low}} = \text{softmax}\left(Q_{\text{Opt}}^{\text{low}\,T} \otimes K_{\text{Opt}}^{\text{low}}\right) \tag{1}$$

$$P_{\text{SAR}}^{\text{low}} = \text{softmax}\left(Q_{\text{SAR}}^{\text{low}\,T} \otimes K_{\text{SAR}}^{\text{low}}\right). \tag{2}$$

Then, the self-attention maps generated from both SAR and optical modalities are input into the cross-attention fusion mechanism. This results in a joint weighted feature map, which can be represented as

$$P_{\text{Fuse}}^{\text{low}} = P_{\text{Opt}}^{\text{low}} \odot P_{\text{SAR}}^{\text{low}} \tag{3}$$

$$AP_{\text{Opt}}^{\text{low}}, AP_{\text{SAR}}^{\text{low}} = \left(P_{\text{Fuse}}^{\text{low}} \odot V_{\text{Opt}}^{\text{low}}\right), \left(P_{\text{Fuse}}^{\text{low}} \odot V_{\text{SAR}}^{\text{low}}\right) \tag{4}$$

$$AP_{Opt-SAR}^{\text{low}} = AP_{\text{Opt}}^{\text{low}} \odot AP_{\text{SAR}}^{\text{low}}. \tag{5}$$

Similarly, the high-level features can be formulated as

$$P_{\text{Fuse}}^{\text{high}} = P_{\text{Opt}}^{\text{high}} \odot P_{\text{SAR}}^{\text{high}} \tag{6}$$

$$AP_{\text{Opt}}^{\text{high}}, AP_{\text{SAR}}^{\text{high}} = \left(P_{\text{Fuse}}^{\text{high}} \odot V_{\text{Opt}}^{\text{high}}\right), \left(P_{\text{Fuse}}^{\text{high}} \odot V_{\text{SAR}}^{\text{high}}\right) \tag{7}$$

$$AP_{Opt-SAR}^{\text{high}} = AP_{\text{Opt}}^{\text{high}} \odot AP_{\text{SAR}}^{\text{high}}. \tag{8}$$

3) Low–High-Level Feature Fusion: To enhance the integration of the extracted joint feature maps, a multiscale feature fusion module is proposed, ensuring the complementarity of low-to-high-level features. The feature fusion mechanism first endeavors to concatenate the joint feature map alongside the cross-modality low/high-level features, expressed as

$$F_{Opt-SAR}^{\text{low}} = \text{concat}\left(F_{\text{SAR}}^{\text{low}}, F_{\text{Opt}}^{\text{low}}, AP_{Opt-SAR}^{\text{low}}\right) \tag{9}$$

$$F_{Opt-SAR}^{\text{high}} = \text{concat}\left(F_{\text{SAR}}^{\text{high}}, F_{\text{Opt}}^{\text{high}}, AP_{Opt-SAR}^{\text{high}}\right). \tag{10}$$

Subsequently, an **ASPP** module is introduced to aggregate contextual information across varying scales from the high-level features. This is particularly beneficial as these high-level features inherently possess a more expansive receptive field. Specifically, for the generated $F_{\text{Opt-SAR}}^{\text{high}}$ via cross-attention, 3×3 convolutions with sampling rates of 6, 12, and 18 are employed to capture features spanning diverse receptive field

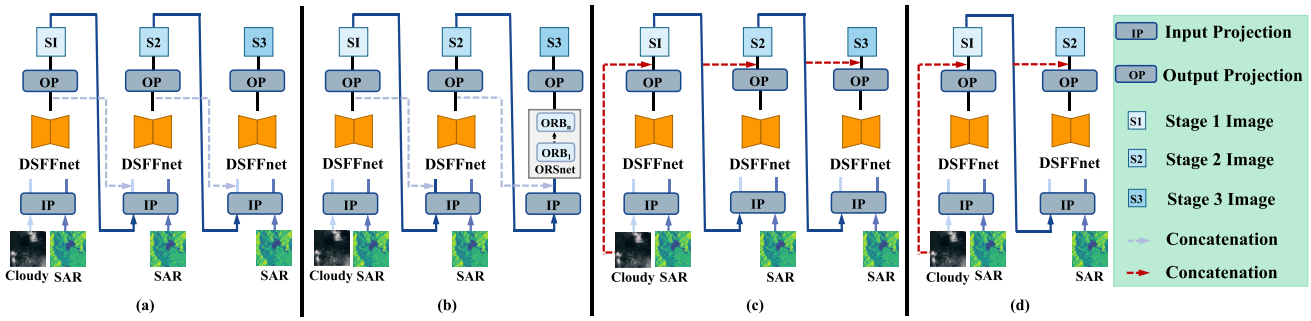


Fig. 3. Workflow of multistage remote sensing image restoration algorithms. (a) Architecture diagram of MPIR1. (b) Architecture diagram of MPIR2. The gray box represents ORSnet. (c) Architecture diagram of MPIR3. (d) Architecture diagram of MPIR4. The blue dashed arrows and red dashed arrows represent concatenation. The solid blue lines represent the transfer of information between two stages and the arrows point to the direction of information transfer.

sizes. Additionally, pooling operations employ a 1×1 convolutional kernel to amalgamate and integrate channel information at each spatial location of the feature map. As depicted in Fig. 1, the feature maps extracted from the four components, possessing identical dimensions, are concatenated. This is followed by feature fusion and dimensionality reduction via a 1-D convolutional layer, culminating in the final representation $F_{\text{Opt-SAR}}^{high_{\text{ASPP}}}$

$$F_{Opt-SAR}^{high_{\text{ASPP}}} = ASPP\left(F_{Opt-SAR}^{\text{high}}\right). \tag{11}$$

Finally, both the high-level and low-level feature maps are concatenated and jointly utilized as input to the decoder

$$F_{Opt-SAR}^{Fuse} = \text{concat}\left(F_{Opt-SAR}^{high_{\text{ASPP}}}, F_{Opt-SAR}^{\text{low}}\right). \tag{12}$$

B. Multistage Progressive Image Restoration

To improve the quality of the restored images, we employ the MPIR strategy, which progressively refines images across multiple stages. The restoration process commences at the first stage using a low-resolution variant of the input image. Subsequently, the output from each stage is channeled as input to its successor, culminating in the acquisition of the final high-resolution image. We have developed four MPIR strategies, each characterized by unique feature connection methodologies, to facilitate high-quality image restoration (as depicted in Fig. 3).

1) MPIR1: In the first MPIR strategy, a three-stage image restoration process is adopted, incorporating intermediate feature connections. The MMCANet, as previously mentioned, functions as the primary feature extraction network, yielding fused cross-multimodal features tailored for restoration endeavors. Notably, features produced by each stage, prior to the terminal convolution, serve as connectors between respective stages. As illustrated by the blue dashed lines, these features are subsequently merged with the ones stemming from the new input postconvolution.

2) MPIR2: Although the MMCANet is adept at assimilating extensive contextual information, there is a risk of local semantic information loss. To address this, our second MPIR strategy substitutes the MMCANet in the third restoration stage with an original-resolution subnetwork (ORSnet) [29]. This approach operates convolutions at the native resolution, ensuring the retention of delicate local features. The structure of the ORB block, depicted in Fig. 4, is primarily built

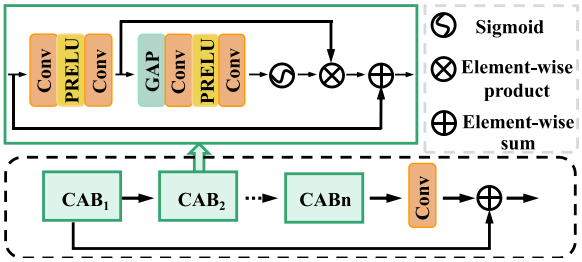


Fig. 4. Detailed diagram of the ORB in ORSNet. The ORB consists of multiple CAB blocks, with the details of a CAB block shown within the green box. GAP stands for global average pooling.

from multiple channel attention blocks (CABs) [30]. Through weighting each channel, the CAB modulates the significance of individual channels across the feature map, thus augmenting the network's feature extraction capability.

3) MPIR3: Inspired by the residual structure [31], which introduces shortcuts to facilitate the flow of information throughout the network, we replace the intermediate connection with skip connections across different restoration stages. This approach not only bolsters the network's robustness but also refines the generated image by amalgamating features from the native resolution data.

4) MPIR4: Generally, traditional progressive restoration algorithms encompass three distinct stages. However, in scenarios where the depth network might amplify noise inherent in the input image or artifacts within the produced image, the result could be suboptimal restoration outcomes. To counteract this, we have streamlined the restoration process in MPIR3 to a two-stage approach. Empirical evaluations underscore the efficacy of this modification.

C. Optimization Objective

Given the predicted image X and the ground truth Y, we utilize the L_1 (average absolute error) as the fundamental error function, which is defined as follows:

$$L_1 = \frac{\|X - Y\|_1}{N} \tag{13}$$

where N is the total number of pixels.

Structural similarity index (SSIM) quantifies the similarity between two images and assesses the quality degradation by measuring the loss of structural information

$$SSIM = \frac{(2\mu_X\mu_Y + C_1)(2\sigma_{XY} + C_2)}{\left(\mu_X^2 + \mu_Y^2 + C_1\right)\left(\sigma_X^2 + \sigma_Y^2 + C_2\right)} \tag{14}$$

where μ, σ, σ_{XY} denote the mean, variance, and covariance of X and Y. $C_1 = 0.01^2$ and $C_2 = 0.03^2$ to avoid zero numerator or denominator.

In image restoration, SSIM can be utilized as a loss function during training to guide the optimization process and enhance the fidelity of the reconstructed images. By minimizing the SSIM loss, the generated images can better preserve structural details, textures, and overall visual similarity with the original images. The mathematical formulation of the SSIM loss function is as follows:

$$L_{\text{SSIM}} = 1 - \frac{1}{N}\sum_{p=1}^{N} SSIM(p) \tag{15}$$

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

To acquire cloudless images featuring sharp boundaries, a custom loss function L_{sum} is derived by taking the sum of the two aforementioned loss functions

$$L_{\text{sum}} = L_1 + L_{\text{SSIM}}. \tag{16}$$

D. Evaluation Metrics

The performance of cloud removal is assessed using four widely used metrics: SSIM, mean absolute error (MAE), spectral angle mapper (SAM), and peak signal-to-noise ratio (PSNR). PSNR assesses the ratio between the maximum achievable power of a signal and the power of the noise impacting the signal, typically expressed in decibels (dB). Elevated PSNR values signify heightened resemblance between the reconstructed image and the original image, implying superior image quality with reduced distortion or noise. MAE quantifies the average pixel-wise disparity between the reconstructed image and the ground truth image, providing a measure of overall reconstruction accuracy by accounting for absolute differences regardless of error direction. A diminished MAE value denotes a smaller average error magnitude. PSNR, along with SSIM, and MAE are employed to evaluate spatial structure restoration, while SAM denotes the degree of spectral information preservation in the restored outcomes.

For the predicted image X and the ground truth Y with pixel values x_i and y_i, the metrics are calculated as given by

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - x_i| \tag{17}$$

$$\text{mse} = \frac{1}{n}\sum_{i=1}^{n} |y_i - x_i|^2 \tag{18}$$

$$\text{PSNR} = 20\log_{10}\left(\frac{1}{\sqrt{\text{mse}}}\right) \tag{19}$$

where n represents the total number of pixels, and mse means mean square error.

SAM treats each pixel's spectrum as a high-dimensional vector and evaluates spectral similarity by computing the angle between two vectors. A smaller angle indicates greater similarity between the spectra

$$\text{SAM} = \arccos\left(\frac{\sum_{i=1}^{n} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \cdot \sum_{i=1}^{n} y_i^2}}\right). \tag{20}$$

IV. Experiment

A. Training Details

The proposed framework is implemented through the PyTorch framework, which is publicly accessible. The training process is set to run for a maximum of 50 epochs to ensure convergence. The learning rate is 1×10^{-4} and is reduced by a factor of 0.1 every 25 epochs. The batch size is 32.

To improve efficiency, all experiments are executed on GPU and leverage a dual-core parallel. Network parameters are optimized using the Adam optimizer [32].

B. Dataset

All experiments are conducted using the SEN12MS-CR [33] dataset, which is a publicly available large-scale dataset for cloud removal in remote sensing. It consists of four sub-datasets corresponding to the seasons: spring, summer, fall, and winter. Taking the spring sub-dataset as an example, each image in the dataset, which has a size of 256×256, is composed of corresponding two-band SAR images, 13-band cloud-free, and cloud-afflicted optical images.

To improve the efficiency of model training, we choose the 42 interested regions (IROs) in spring data. The SEN12MS-CR dataset is not officially divided into dataset, validation set, and test set. In order to prevent overfitting in model training, we choose to divide the dataset, validation set, and test set in the ratio of 20:1:1, so that the training set has sufficient data volume. Specifically, we select two IROs with relatively evenly distributed cloud cover at different levels as a test set. The other 38 IROs as the training set, while two additional IROs are designated as the validation set. The training set consists of 25 521 images, the validation set includes 1244 images, and the test set has 1453 images. The images are all resized to 128×128 pixels. In addition, following the data processing approach of [13], we crop the first channel data of SAR images to [-25, 0], the second channel data to [-32.5, 0], and finally scaled them to [0, 1]. For cloud-free optical images and cloudy images, we only select the RGB channels, cropped them to [0, 10 000], and normalized them to [0, 1]. This data processing approach is used for all experiments presented in this article. Moreover, in order to prevent overfitting during testing, based on the original spring test set, we select datasets from other seasons as the test set to evaluate the generalization ability of the model. Specifically, we randomly select 951 images from the summer sub-dataset, 1514 images from the fall sub-dataset, and 954 images from the winter sub-dataset.

C. Comparative Baselines

In this article, we compare our proposed method with several typical cloud removal algorithms.

1) *McGAN [34]:* An approach that uses the joint data of SAR and optical images as input and employs the U-Net as a generator to restore cloud-free images from cloudy ones.
2) *SpA GAN [35]:* A method that employs the spatial attention [36] as the generator, which only utilizes cloudy

TABLE I
PERFORMANCE OF OURS METHOD ON THE SEN12MS-CR DATASET

Season	PSNR↑	SSIM↑	MAE↓	SAM↓
Spring	39.3497	0.9601	0.0088	2.9003
Summer	30.8009	0.9072	0.0265	2.7979
Fall	36.2944	0.9408	0.0196	5.7017
Winter	33.8424	0.9371	0.0173	4.5639

images as input and does not incorporate auxiliary SAR images.

3) *DSen2-CR [12]:* A residual structure based on the modification of the Deep Sentinel-2 ResNet [37] which takes the concatenated optical and SAR images as input.
4) *GLF-CR [13]:* It stacks multiple local modules for information interaction between the two modalities, which significantly improves the effectiveness of image restoration.
5) *Uformer [19]:* A network derived from a U-shaped transformer-based network [38], which also takes the concatenated data of SAR and optical images as input.

D. Experiments of All Methods on SEN12MS-CR

To rigorously evaluate the generalization ability of our proposed method, we conduct both in-domain and out-of-domain experiments on all baseline techniques. The in-domain evaluation focuses on analyzing the removal performance under conditions similar to the training dataset. In contrast, the out-of-domain test measures the effectiveness of the removal process on unseen data, especially in different seasons such as summer, autumn, and winter. Table I summarizes the quantitative results of our method in-domain and out-of-domain, which demonstrates the feasibility and generalization of the proposed method. Tables II and III show the quantitative comparison results of our method with all baselines, further verifying that the proposed method is a reliable solution for cloud removal. For all experimental results, a more detailed analysis will be given in Section V.

E. Analysis of MPIR for Could Removal

Similarly, we conduct both in-domain and out-of-domain experiments to assess the effectiveness of the proposed MPIR strategy in comparison to the standalone MMCANet.

1) In-Domain Testing: Fig. 5 presents a visual comparison of the MPIR variants against the standalone MMCANet. We can observe that all MPIR variants surpass the performance of the MMCANet baseline. The red boxes emphasize the specific regions chosen for a detailed comparison. We choose a range of terrains with different cloud cover intensities to investigate the proficiency of progressive algorithms in conserving the original regions and maintaining edge details. A scrutiny reveals that all four algorithms yield commendable results in cloud removal without introducing discernible artifacts. Evaluating color fidelity, MPIR1, MPIR2, and MPIR4 display a color distribution that closely aligns with the original image. Conversely, MPIR3 lags behind in this metric. When it comes to edge preservation, MPIR1, MPIR2, and MPIR4 hold a distinct edge over MMCANet.

Furthermore, we note that MPIR1's cloud removal efficacy diminishes significantly in regions with dense cloud cover, leading to lingering pseudo-shadows, particularly evident in the red regions of the fourth row. The granularity of recovery in the cloud-covered areas also leaves room for improvement. Conversely, in cloud-free zones, MPIR2 consistently outperforms MMCANet in preserving inherent data. This superior preservation is largely credited to the deliberate architecture of MPIR2's concluding phase, designed explicitly for high-fidelity resolution maintenance. Nonetheless, in terms of rejuvenating areas obscured by clouds, while the global land contour is discernibly reconstructed, the intricate masonry patterns within remain somewhat indistinct.

Additionally, we observe that the cloud removal efficacy of the MPIR3 strategy is inferior to that of the single-stage MMCANet irrespective of the cloud density. One possible explanation for MPIR3's mediocre performance in image restoration is the excessive depth of the network, which may lead to a loss of intricate details and local information during propagation. To address this issue, we have streamlined the depth of the MPIR3 and proposed a two-tiered architecture named MPIR4. This updated model, MPIR4, demonstrates enhanced texture and structural fidelity in the reconstructed images. In areas devoid of cloud interference, the restored results closely resemble the original image.

To evaluate the cloud removal performance quantitatively, we utilized four distinct metrics on the spring dataset. Detailed results are tabulated in Table IV. While MMCANet serves as a benchmark, both MPIR4 and MPIR2 demonstrate marked enhancements in certain metrics, underlining the efficacy of these multistage algorithms for cloud removal tasks. The suboptimal values for MPIR1 can be ascribed to its constraints in addressing cloud-covered areas. As for MPIR3, its diminished metric scores can largely be attributed to the pronounced information loss within the deep network structure.

2) Out-of-Domain Testing: Similarly, we also conduct experiments under out-of-domain settings to assess the robustness of MPIR algorithms. The quantitative results are presented in Table IV, and we can observe that MPIR4 and MPIR2 demonstrate stable performance across all datasets, exhibiting strong robustness. In contrast, the performance of MPIR1 and MPIR3 displays inconsistency depending on the dataset, hinting at a reduced capacity for generalization.

F. Ablation Analysis

1) Major Components: To evaluate the impact of the key components in the proposed MMCANet, we perform ablation studies in an in-domain setting. We conduct the effects of each component (i.e., SAR integration, cross-attention module, and ASPP) and their various combinations. Ablation studies results are presented in Table V, indicating a positive contribution from each component within the MMCANet framework toward the final outcomes. Since the dual-stream network configuration becomes inapplicable when only cloudy optical images are employed as inputs, we modify the input by concatenating the channels of SAR and optical images.

TABLE II

IN-DOMAIN PERFORMANCE COMPARISON. MMCANET REPRESENTS THE USE OF 3 × 3 CONVOLUTION IN DWCONV WITHIN MFF, WHILE MMCANET$_L$ REPRESENTS THE USE OF 5 × 5 CONVOLUTION IN DWCONV WITHIN MFF

Model	Network		Input		Spring			
	Single	Dual	SAR	Optical	PSNR (dB) ↑	SSIM ↑	MAE ↓	SAM (°) ↓
McGAN	✓		✓	✓	32.8749	0.9029	0.0202	6.4730
SpA GAN	✓		×	✓	31.2022	0.8434	0.0248	5.7881
DSen2-CR	✓		✓	✓	35.6723	0.9383	0.0119	3.9167
GLF-CR		✓	✓	✓	36.4810	0.9158	0.0111	4.4319
Uformer	✓		✓	✓	35.8721	0.9462	0.0132	4.3070
MMCANet (Ours)		✓	✓	✓	**39.3497**	**0.9601**	**0.0088**	**2.9003**
$MMCANet_l$ **(Ours)**		✓	✓	✓	39.8871	0.9672	0.0081	2.9884

TABLE III

OUT-OF-DOMAIN PERFORMANCE COMPARISON. MMCANET REPRESENTS THE USE OF 3 × 3 CONVOLUTION IN DWCONV WITHIN MFF, WHILE MMCANET$_L$ REPRESENTS THE USE OF 5 × 5 CONVOLUTION IN DWCONV WITHIN MFF

Model	Summer				Fall				Winter			
	PSNR↑	SSIM↑	MAE↓	SAM↓	PSNR↑	SSIM↑	MAE↓	SAM↓	PSNR↑	SSIM↑	MAE↓	SAM↓
McGAN	28.0349	0.8762	0.0299	4.6122	30.5847	0.8771	0.0286	6.3197	31.6399	0.9177	0.0219	4.9394
SpA GAN	28.7614	0.7710	0.0256	6.5527	30.3656	0.8381	0.0281	**4.7607**	28.9497	0.8561	0.0333	4.8489
DSen2-CR	30.8197	0.8955	0.0240	3.6090	35.8479	0.9392	**0.0140**	5.7637	33.0852	0.9330	0.0176	**3.9185**
GLF-CR	30.6023	0.8846	0.0261	3.3700	35.4592	0.9329	0.0148	5.7844	33.6379	0.9326	0.0178	4.4617
Uformer	**31.3746**	0.9067	**0.0230**	2.8234	34.2890	0.9386	0.0174	5.3681	33.7890	**0.9386**	0.0174	5.3681
MMCANet (Ours)	30.8009	**0.9072**	0.0265	**2.7979**	**36.2944**	**0.9408**	0.0196	5.7017	**33.8424**	0.9371	**0.0173**	4.5639
$MMCANet_l$ **(Ours)**	31.5596	0.9110	0.0233	2.4492	36.1682	0.9515	0.0140	5.8007	33.0124	0.9425	0.0197	4.7185

The numeric results demonstrate the pivotal role of SAR modality integration in providing extra information for the cloud removal task. Furthermore, the pronounced boost in SSIM underscores the capacity of cross-attention to better harness the texture details from SAR images, underscoring the significance of merging SAR and optical imagery. Additionally, implementing ASPP has invariably led to improvements across all evaluative metrics. This insight suggests that elevating the network's feature extraction prowess in a constructive manner can furnish a wealth of contextual details, a factor paramount for amplifying network efficacy.

2) Loss Components: To assess the beneficial impact of the SSIM loss in the task of cloud removal, we contrasted the performance of MMCANet on the spring test dataset, considering both the L_1 loss and L_{sum} loss as independent variables. The findings are detailed in Table VI. The marked enhancement in the assessment metrics underscores the importance of choosing the right loss function, which can refine the network's effectiveness and lead to pronounced performance improvements.

V. DISCUSSION

In this section, we will analyze in detail the performance of the proposed method compared with other baselines for cloud removal.

To evaluate the performance of various cloud removal algorithms, we select six scenarios with cloud coverage ranging from 0%–10%, 10%–20%, 20%–40%, 40%–60%, 60%–80%, and 80%–100%. Fig. 6 presents the visual results of various algorithms for cloud removal. As shown in Fig. 6, our method achieves higher image fidelity than others, meaning that the generated images closely match the color distribution of the ground truth. McGAN and GLF-CR exhibit compromised color fidelity. This deficiency might arise from the models' inability to accurately capture the color mapping relationship between input and output images, leading to a notable degradation in the visual quality of their results.

Furthermore, compared to alternative networks, our method showcases superior image restoration, preserving intricate details under both thick and thin cloud conditions. McGAN and SpA GAN both utilize analogous generator architectures. However, McGAN surpasses SpA GAN in texture restoration, owing to the additional texture information furnished by SAR. However, when the cloud coverage exceeds 20%, the McGAN also loses its cloud removal capability. DSen2-CR exhibits limited capability in cloud and cloud shadow removal, as undesirable artifacts are still noticeable in the resulting images. Specifically, under thick clouds that obstruct the line of sight, the generated images can only retain minimal useful information. For GLF-CR, despite its low image fidelity leading to lower image quality, its ability to restore image details can still be observed, as in the scenario in the fourth row. Uformer excels in capturing both local and global dependencies crucial for image restoration, thus markedly improving cloud removal efficacy. While it adeptly manages thin clouds without leaving visible artifacts, its proficiency dwindles when confronted with regions blanketed by dense clouds.

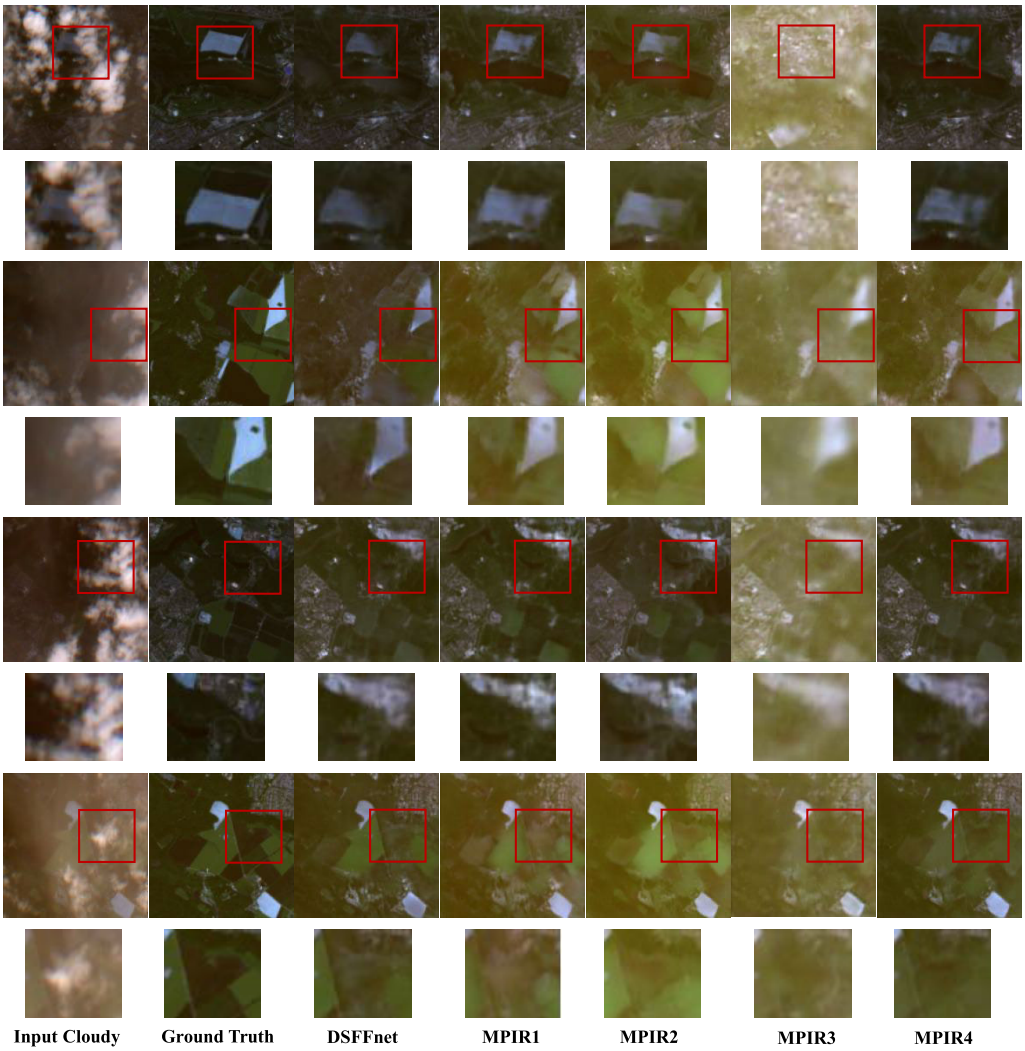


Fig. 5. Visual comparison of MPIR variants and MMCANet for cloud removal. The second row shows the local enlargement of the corresponding regions outlined in red boxes.

TABLE IV

PERFORMANCE COMPARISON OF MPIR VARIANTS. THE BOLD AND ITALIC ENTRIES INDICATE THE OPTIMAL AND SUBOPTIMAL RESULTS, RESPECTIVELY

Model	Spring				Summer				Fall				Winter			
	PSNR↑	SSIM↑	MAE↓	SAM↓	PSNR↑	SSIM↑	MAE↓	SAM↓	PSNR↑	SSIM↑	MAE↓	SAM↓	PSNR↑	SSIM↑	MAE↓	SAM↓
MPIR1	39.1921	0.9542	0.0085	3.0287	31.1029	0.9037	0.0252	**2.9108**	**36.1600**	**0.9414**	**0.0143**	5.6558	33.5852	0.9322	0.0182	4.8248
MPIR2	*39.3795*	*0.9586*	*0.0084*	**2.6557**	**32.1404**	**0.9058**	**0.0220**	3.0500	*36.0776*	*0.9413*	*0.0145*	5.6910	*33.9568*	0.9360	*0.0170*	*4.7920*
MPIR3	38.7171	0.9494	0.0087	3.3434	29.7002	0.8942	0.0302	3.1872	36.0575	0.9393	0.0147	**5.3044**	33.8443	*0.9363*	0.0174	5.0112
MPIR4	**39.7558**	**0.9618**	**0.0080**	*3.0751*	*31.1654*	0.9023	*0.0245*	*2.9989*	35.9628	0.9384	0.0146	*5.6198*	**34.2664**	**0.9411**	**0.0169**	**4.3642**

TABLE V

ABLATION STUDY OF MMCANET OF SAR, CROSS-ATTENTION, AND ASPP

SAR	Cross-Attention	ASPP	PSNR↑	SSIM↑	MAE↓	SAM↓
×	×	×	38.0459	0.9487	0.0097	3.8811
✓	×	×	38.6841	0.9501	0.0091	3.5616
✓	✓	×	38.7398	0.9576	0.0091	3.4150
✓	✓	✓	**39.3497**	**0.9601**	**0.0088**	**2.9003**

TABLE VI

ABLATION EXPERIMENT OF SINGLE-STAGE NETWORK WITH DIFFERENT LOSSES

L_1	L_{SSIM}	PSNR↑	SSIM↑	MAE↓	SAM↓
✓	×	37.5102	0.9206	0.0103	3.4713
✓	✓	**39.3497**	**0.9601**	**0.0088**	**2.9003**

In terms of information preservation in cloud-free regions, our approach consistently outperforms other methods. The McGAN and SpA GAN tend to preserve a blurred texture in these regions while DSen2-CR stands out with its impeccable performance in maintaining clarity. This success can be attributed to its unique residual structure and the

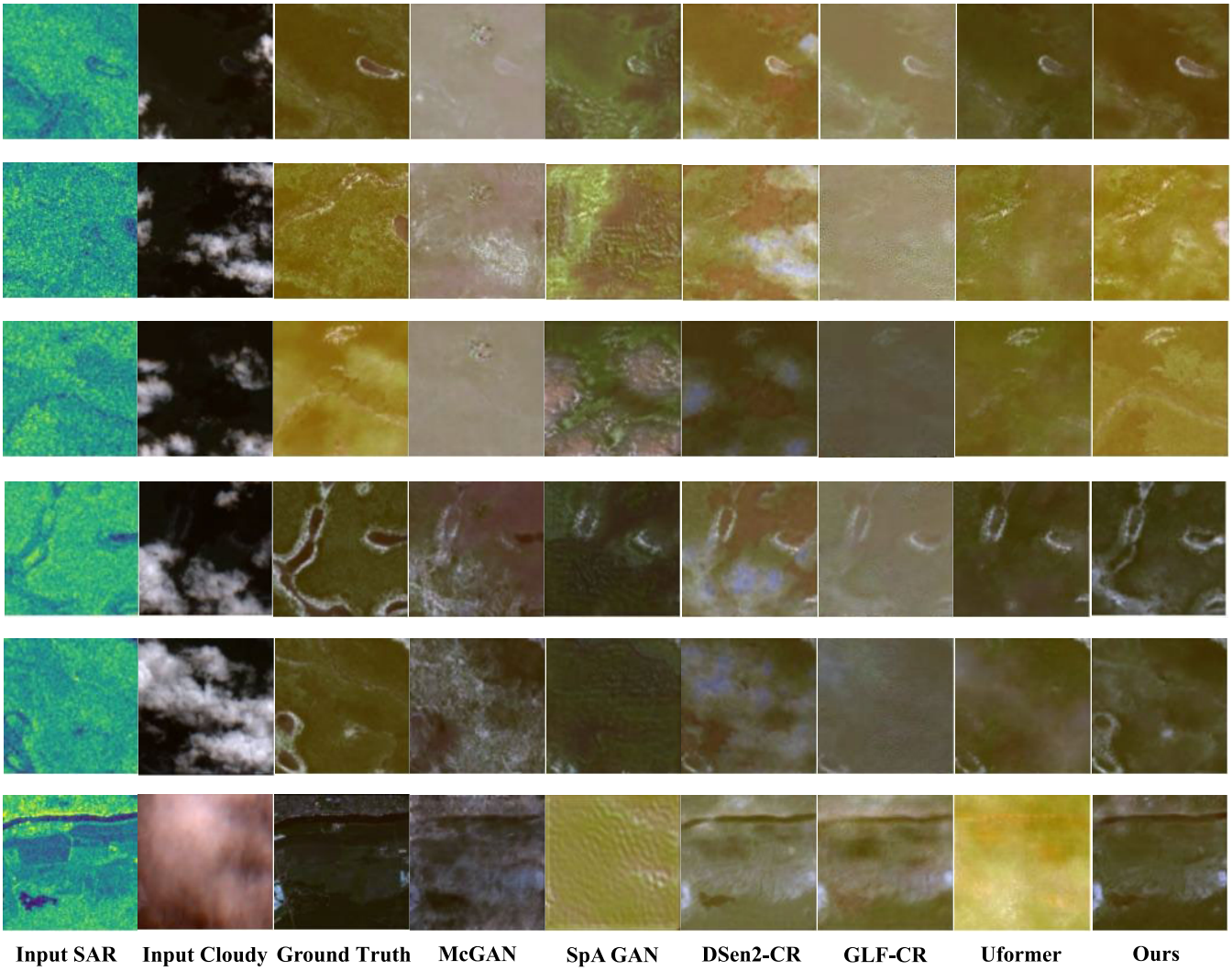


Fig. 6. Comparison results of cloud removal under different cloud and cloud shadow coverage scenes. (Top to bottom) Cloud shadow coverage ranging from 0%–10%, 10%–20%, 20%–40%, 40%–60%, 60%–80%, and 80%–100%, respectively.

cloud-adaptive regularized loss function. Similarly, Uformer also excels in safeguarding the integrity of data in cloud-free zones.

In order to examine the ability of the network to recover details in complex scenarios, we select different complexity-building coverage scenes (as shown in Fig. 7). We can observe that our proposed approach present decent image restoration ability regarding the details of complex buildings, even in the presence of fog or thin/thick clouds. The SpA GAN experiences difficulty in preserving details within complex scenes, largely attributed to the lack of SAR information. In contrast, McGAN, DSen2-CR, and Uformer use the fusion method that combines SAR and optical image channels directly, which produce blurred and less detailed images due to the heterogeneity of multimodal data. While GLF-CR occasionally yields blurred images, it remarkably retains substantial edge details.

Table II showcases the quantitative results obtained from the spring dataset. These metric outcomes align well with our prior visual analysis. Specifically, McGAN, SpA GAN, and GLF-CR display elevated SAM values, pointing to notable spectral drifts and color anomalies. Both McGAN and SpA GAN-generated images exhibit pronounced deviations from the authentic data, translating to reduced likeness. Meanwhile, DSen2-CR demonstrates superior data retention in regions devoid of clouds, leading to commendable quantitative scores. However, due to the limited capabilities of GLF-CR and Uformer in reconstructing images affected by dense clouds, they record a heightened MAE and a diminished SSIM in comparison to our proposed method.

It is imperative to highlight that the disparities in cloud cover distribution across these datasets might lead to variability in the results. For instance, the winter dataset might exhibit a larger count of cloud-free images. As depicted in Table III, elevated SSIM and PSNR values, combined with diminished MAE and SAM metrics, underscore the superior robustness and generalizability of the proposed model. Furthermore, although the SNR of input SAR images varies (even within the same dataset), the performance of the proposed method remains stable. The key factor is the extraction of edge features from the SAR images, as the proposed network performs effectively when the SNR exceeds 20 dB. Based on existing works, it can be assured that this method applies to most spaceborne SAR images, not only from public datasets

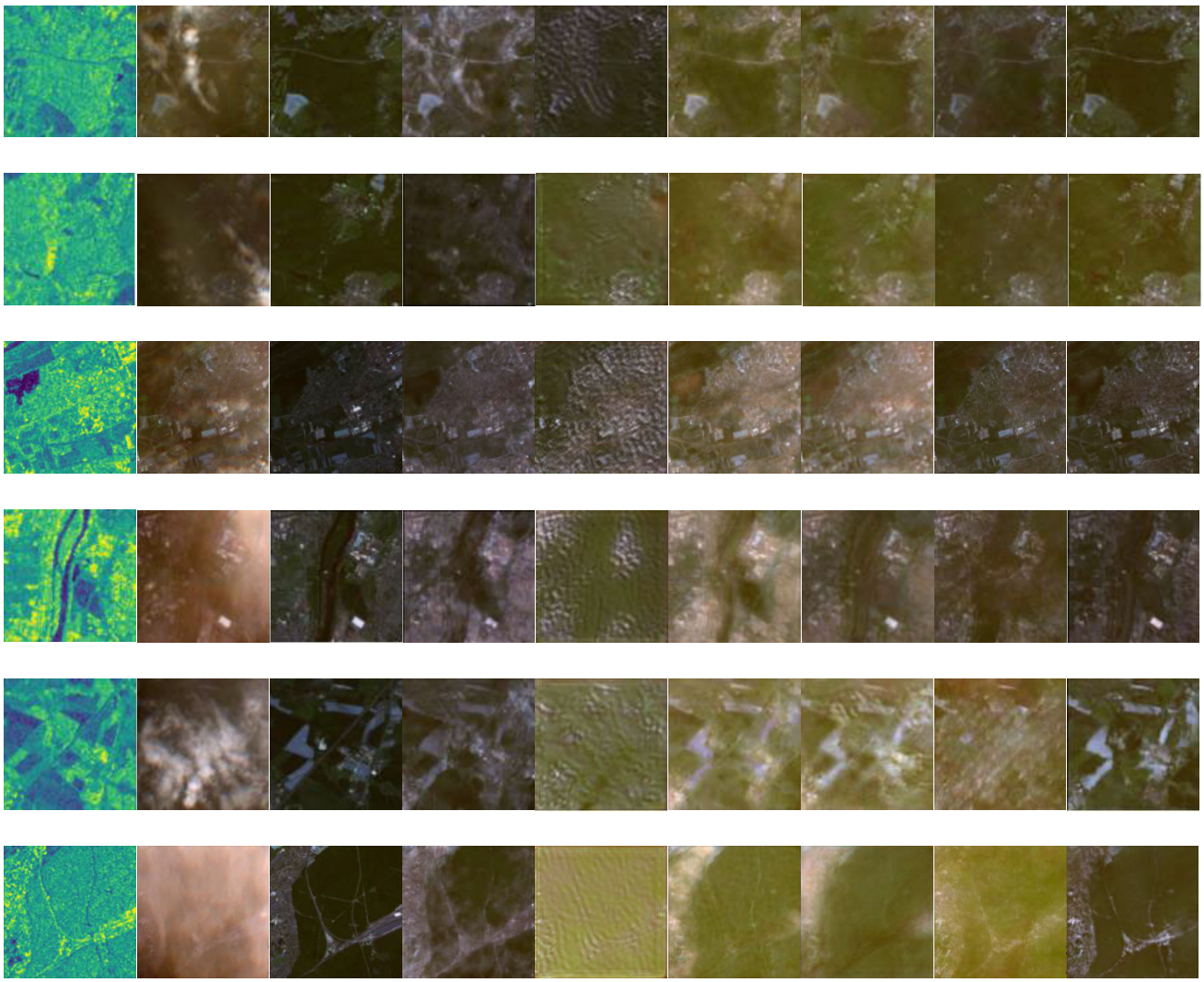


Fig. 7. Comparison results of cloud removal under different building coverage scenes. From left to right, the columns represent SAR, cloudy optical image, ground truth, and comparison results of McGAN, SpA GAN, DSen2-CR, GLF-CR, Uformer, and MMCANet.

but also from open-access data produced by on-orbit satellites, such as Sentinel-1.

Practical Usage of MMCANet: The computational complexity of single-stage network MMCANet is comparatively lower, yet its performance is marginally inferior to that of certain multistage networks. The computational complexity of multistage networks increases with stage expansion to provide better restoration performance. Consequently, in practical applications, the selection of an appropriate cloud removal network should be tailored to the specific remote sensing image application scenarios. For instance, in the case of rapid-onset and extensive flood disasters, there is a heightened demand for the real-time capabilities of cloud removal technologies, making the low-computational single-stage network a more suitable choice. For applications that require coarse image restoration quality, such as some agricultural monitoring, which often involves large-scale areas and does not require high spatial detail, a single-stage network is suitable due to its lower computational complexity. However, for applications like urban planning and management, which require detailed and fine-grained restoration images to capture subtle features and variations, multistage restoration is necessary. For future work, we plan to implement techniques such as model pruning to reduce the parameter count of the multistage network, thereby minimizing computational resource requirements while maintaining high performance.

VI. Conclusion

In this article, we propose a single-stage dual-stream feature fusion network MMCANet along with multistage cloud removal algorithms to address the cloud removal task. A dual-stream network is designed to allow the incorporation of the SAR modality and the optical modality, which provide more fine-grained texture information for image restoration. Subsequently, a cross-attention mechanism is proposed to explore the cross-modal fusion between the high-dimensional features of optical and SAR, enabling the interaction between high-order and low-order features for information complementarity. Furthermore, we utilize the ASPP module in the interacted high-order features to extract multiscale features

and enrich the extracted semantic information to capture the cloud area around the target, significantly improving the ability of cloud cover removal at different levels. The performance of the proposed single-stage network surpasses the state-of-the-art approaches. To further improve the performance, we explore multistage image restoration algorithms. We devise four distinct multistage architectures tailored for remote sensing image restoration and conduct a thorough analysis of their strengths and weaknesses. Experimental results validate that two of our devised progressive algorithms yield notable performance enhancements. In the future, we will work on developing lightweight cloud removal models to enhance the application of the algorithm in edge devices or actual scenarios.

REFERENCES

[1] W. Diao, X. Sun, X. Zheng, F. Dou, H. Wang, and K. Fu, "Efficient saliency-based object detection in remote sensing images using deep belief networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 2, pp. 137–141, Feb. 2016.

[2] N. Kussul, M. Lavreniuk, S. Skakun, and A. Shelestov, "Deep learning classification of land cover and crop types using remote sensing data," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 5, pp. 778–782, 2017.

[3] B. Adriano et al., "Learning from multimodal and multitemporal Earth observation data for building damage mapping," *ISPRS J. Photogramm. Remote Sens.*, vol. 175, pp. 132–143, May 2021.

[4] M. D. King, S. Platnick, W. P. Menzel, S. A. Ackerman, and P. A. Hubanks, "Spatial and temporal distribution of clouds observed by MODIS onboard the Terra and Aqua satellites," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 7, pp. 3826–3852, Jul. 2013.

[5] V. Sarukkai, A. Jain, B. Uzkent, and S. Ermon, "Cloud removal from satellite images using spatiotemporal generator networks," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, Mar. 2020, pp. 1796–1805.

[6] C. Yu, L. Chen, L. Su, M. Fan, and S. Li, "Kriging interpolation method and its application in retrieval of MODIS aerosol optical depth," in *Proc. 19th Int. Conf. Geoinformatics*, Jun. 2011, pp. 1–6.

[7] M. Xia and K. Jia, "Reconstructing missing information of remote sensing data contaminated by large and thick clouds based on an improved multitemporal dictionary learning method," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5605914.

[8] P. Ebel, M. Schmitt, and X. X. Zhu, "Cloud removal in unpaired Sentinel-2 imagery using cycle-consistent GAN and SAR-optical data fusion," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Sep. 2020, pp. 2065–2068.

[9] W. He and N. Yokoya, "Multi-temporal Sentinel-1 and -2 data fusion for optical image simulation," *ISPRS Int. J. Geo-Information*, vol. 7, no. 10, p. 389, Sep. 2018.

[10] J. D. Bermudez, P. N. Happ, D. A. B. Oliveira, and R. Q. Feitosa, "SAR to optical image synthesis for cloud removal with generative adversarial networks," *ISPRS Ann. Photogramm., Remote Sens. Spatial Inf. Sci.*, vols. IV–1, pp. 5–11, Sep. 2018.

[11] C. Grohnfeldt, M. Schmitt, and X. Zhu, "A conditional generative adversarial network to fuse SAR and multispectral optical data for cloud removal from Sentinel-2 images," in *Proc. IGARSS IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2018, pp. 1726–1729.

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

[13] F. Xu et al., "GLF-CR: SAR-enhanced cloud removal with global–local fusion," *ISPRS J. Photogramm. Remote Sens.*, vol. 192, pp. 268–278, Oct. 2022.

[14] Y. Wang, B. Liang, M. Ding, and J. Li, "Dense semantic labeling with Atrous spatial pyramid pooling and decoder for high-resolution remote sensing imagery," *Remote Sens.*, vol. 11, no. 1, p. 20, Dec. 2018.

[15] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2017.

[16] K. Enomoto et al., "Filmy cloud removal on satellite imagery with multispectral conditional generative adversarial nets," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, Jul. 2017, pp. 48–56.

[17] H. Pan, "Cloud removal for remote sensing imagery via spatial attention generative adversarial network," 2020, *arXiv:2009.13015*.

[18] J. Gao, Q. Yuan, J. Li, H. Zhang, and X. Su, "Cloud removal with fusion of high resolution optical and SAR images using generative adversarial networks," *Remote Sens.*, vol. 12, no. 1, p. 191, Jan. 2020.

[19] S. Han, J. Wang, and S. Zhang, "Former-CR: A transformer-based thick cloud removal method with optical and SAR imagery," *Remote Sens.*, vol. 15, no. 5, p. 1196, Feb. 2023.

[20] Y. A. Farha and J. Gall, "MS-TCN: Multi-stage temporal convolutional network for action segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3575–3584.

[21] B. Filtjens, B. Vanrumste, and P. Slaets, "Skeleton-based action segmentation with multi-stage spatial–temporal graph convolutional neural networks," *IEEE Trans. Emerg. Topics Comput.*, vol. 12, no. 1, pp. 202–212, Dec. 2022.

[22] Z. Shou, D. Wang, and S.-F. Chang, "Temporal action localization in untrimmed videos via multi-stage CNNs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1049–1058.

[23] Y. Bai, M. Liu, C. Yao, C. Lin, and Y. Zhao, "MSPNet: Multi-stage progressive network for image denoising," *Neurocomputing*, vol. 517, pp. 71–80, Jan. 2023.

[24] L. Yan, M. Zhao, S. Liu, S. Shi, and J. Chen, "Cascaded transformer U-net for image restoration," *Signal Process.*, vol. 206, May 2023, Art. no. 108902.

[25] C. M. Manu, "MSDNet: A novel multi-stage progressive image dehazing network," in *Proc. 12th Indian Conf. Comput. Vis., Graph. Image Process.*, Dec. 2021, pp. 1–9.

[26] Y. Zhang et al., "KBNet: Kernel basis network for image restoration," 2023, *arXiv:2303.02881*.

[27] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1251–1258.

[28] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2018, pp. 7132–7141.

[29] S. W. Zamir et al., "Multi-stage progressive image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 14821–14831.

[30] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 286–301.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning," *Image Recognit.*, vol. 7, no. 4, pp. 327–336, 2015.

[32] T. Dozat, "Incorporating Nesterov momentum into Adam," in *Proc. ICLR*, 2016, pp. 1–21.

[33] P. Ebel, A. Meraner, M. Schmitt, and X. X. Zhu, "Multisensor data fusion for cloud removal in global and all-season Sentinel-2 imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5866–5878, Jul. 2021.

[34] Y. Mroueh, T. Sercu, and V. Goel, "McGan: Mean and covariance feature matching GAN," in *Proc. Int. Conf. Mach. Learn.*, Jan. 2017, pp. 2527–2535.

[35] H. Emami, M. M. Aliabadi, M. Dong, and R. B. Chinnam, "SPA-GAN: Spatial attention GAN for image-to-image translation," *IEEE Trans. Multimedia*, vol. 23, pp. 391–401, 2021.

[36] T. Wang, X. Yang, K. Xu, S. Chen, Q. Zhang, and R. W. H. Lau, "Spatial attentive single-image deraining with a high quality real rain dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12270–12279.

[37] C. Lanaras, J. Bioucas-Dias, S. Galliani, E. Baltsavias, and K. Schindler, "Super-resolution of Sentinel-2 images: Learning a globally applicable deep neural network," *ISPRS J. Photogramm. Remote Sens.*, vol. 146, pp. 305–319, Dec. 2018.

[38] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, and H. Li, "Uformer: A general U-shaped transformer for image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17683–17693.

Yejian Zhou (Member, IEEE) was born in Zhejiang, China, in 1993. He received the B.S. degree in electronic engineering and the Ph.D. degree in signal processing from Xidian University, Xi'an, China, in 2015 and 2020, respectively.

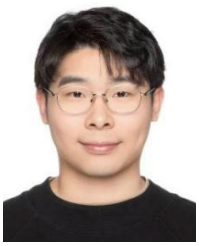
He was a Visiting Ph.D. Student with the Department of Urban Planning and Environment, KTH Royal Institute of Technology, Stockholm, Sweden, from September 2019 to August 2020. He is currently an Associate Professor with the College of Information Engineering, Zhejiang University of Technology, Hangzhou, China. His research interests include ISAR imaging and image interpretation.

Jiahui Suo was born in Henan in 1998. She received the bachelor's degree in engineering from North China University of Water Resources and Electric Power, Zhengzhou, China, in 2021 and the Master of Engineering degree from Zhejiang University of Technology, Hangzhou, China, in 2024.

Her research interests include cloud removal in synthetic aperture radar (SAR) images.

Yachen Wang was born in Shandong in 2000. He received the bachelor's degree in engineering from Wuhan University of Science and Technology, Wuhan, China, in 2022. He is currently pursuing a master's degree with Zhejiang University of Technology, Hangzhou, China.

His research interests include semantic segmentation and height estimation.



Jie Su received the B.S. degree in computer science and technology from China Jiliang University, Hangzhou, China, in 2017, and the M.S. degree (Hons.) in data analytics from the University of Southampton, Southampton, U.K., in 2018, and the Ph.D. degree from Newcastle University, Newcastle upon Tyne, U.K., in 2023.

He is currently an Assistant Professor at the Institute of Cyberspace Security and College of Information Engineering, Zhejiang University of Technology, Hangzhou. His research interests include deep learning, signal processing, and IoT security.

Wen Xiao (Member, IEEE) received the B.S. degree in geodesy and geomatics from Wuhan University, Wuhan, China, in 2010, the M.S. degree (cum laude) in geoinformatics from the Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, Enschede, The Netherlands, in 2012, and the Ph.D. degree in geoinformation science and technology from the National Institute of Geographical Information and Forestry (IGN), Université Paris-EST, Paris, France, in 2016.

He has been a Lecturer and an University Research Fellow with Newcastle University, Newcastle upon Tyne, U.K. He is currently a Professor with China University of Geosciences, Wuhan. His research interests include 3-D mapping, laser scanning, photogrammetric computer vision, and their applications in smart cities and digital twins.

Zhen Hong (Member, IEEE) received the joint B.S. degree from Zhejiang University of Technology, Hangzhou, China, and the University of Tasmania, Hobart, TAS, Australia, in 2006, and the Ph.D. degree from Zhejiang University of Technology, Hangzhou, China, in 2012.

He was a Research Scholar at the CAP Research Group, School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA, from 2016 to 2018. He is currently a Full Professor at the Institute of Cyberspace Security and the College of Information Engineering, Zhejiang University of Technology. His research interests include the Internet of Things, cyberspace security, and data analytics.

Dr. Hong is a member of ACM and a Senior Member of CCF and CAA. He received the first Zhejiang Provincial Young Scientists Title in 2013 and Zhejiang Provincial New Century 151 Talent Project in 2014. He also received Zhejiang Provincial Science Fund for Distinguished Young Scholars in 2023.

Rajiv Ranjan (Fellow, IEEE) is the University Chair Professor of the Internet of Things research with the School of Computing, Newcastle University, Newcastle upon Tyne, U.K. He is the Director of the with School of Computing, Networked and Ubiquitous Systems Engineering (NUSE) Group, Newcastle upon Tyne, jointly with Dr. Graham Morgan. He is also the Academic Director of the School of Computing, Newcastle upon Tyne, and the Research Director of Newcastle Urban Observatory, Newcastle upon Tyne. He is also the Founding Director of the International Centre (U.K.–Australia) on the Internet of Energy (IoE) funded by EPSRC. He is an internationally established Scientist in the area of distributed systems (having published over 250 scientific papers out of which about 50 papers in IEEE/ACM TRANSACTIONS journals).

Dr. Ranjan is a fellow of the Academia Europaea and Asia Pacific Artificial Intelligence Association.

Lizhe Wang (Fellow, IEEE) received the B.E. and M.E. degrees from Tsinghua University, Beijing, China, in 1998 and 2001, respectively, and the D.E. degree (magna cum laude) from the University of Karlsruhe, Karlsruhe, Germany, in 2007.

He is currently a ChuTian Chair Professor with the School of Computer Science, China University of Geosciences, Beijing. He is a Professor with the Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing. His research interests include high-performance computing, e-Science, and remote sensing image processing.

Dr. Wang is a fellow of IET and the British Computer Society. He serves as an Associate Editor for IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, IEEE TRANSACTIONS ON CLOUD COMPUTING, and IEEE TRANSACTIONS ON SUSTAINABLE COMPUTING.

Zhenyu Wen (Senior Member, IEEE) is currently the Tenure-Tracked Professor with the Institute of Cyberspace Security and College of Information Engineering, Zhejiang University of Technology, Hangzhou, China. His research interests include the IoT, crowd sources, AI systems, and cloud computing. For his contributions to the area of scalable data management for the Internet of Things.

Dr. Wen was awarded the IEEE TCSC Award for Excellence in Scalable Computing (Early Career Researchers) in 2020.