# CERMF-Net: A SAR-Optical Feature Fusion for Cloud Elimination From Sentinel-2 Imagery Using Residual Multiscale Dilated Network

Jayakrishnan Anandakrishnan, Venkatesan M Sundaram, and Prabhavathy Paneer

***Abstract*—Satellite-based Earth observation activities, such as urban and agricultural land monitoring, change detection, and disaster management, are constrained by adequate spatial and temporal ground observations. The presence of aerosols and clouds usually distorts quality ground optical observations and reduces the temporal resolution, which degrades the learning and extraction of valuable information. The uncertainty in the occurrence of clouds in the Earth's atmosphere and the possible land changes in subsequent temporal visits is the major challenge in cloud-free reconstruction problems. Advancements in deep learning enabled learning from multisensory inputs, and cloud removal problem seek helps from auxiliary information for better reconstruction. This research introduces a synthetic aperture radar (SAR) guided feature Fusion for Cloud Elimination from Sentinel-2 multispectral imagery using Residual Multiscale dilated Network (CERMF-Net). The proposed CERMF-Net fuses SAR with Sentinel-2 optical data and learn spatial–temporal dependencies and physical–geometrical properties for effective cloud removal. The generalizability and robustness of CERMF-Net are tested against the SEN12MS-CR dataset, a global real cloud-removal dataset. The CERMF-Net displayed superior performance in comparison with the state-of-the-art techniques.**

***Index Terms*—Cloud removal, data fusion, multiscale convolutional neural network (CNN), Sentinel-2, synthetic aperture radar (SAR).**

## I. Introduction

EARTH observation applications heavily depend on the spatial–temporal data from space-borne satellites. Random clouds in the atmosphere significantly impede optical observations from the satellites because clouds are opaque to the optical spectrum. There is no guarantee that the environment will be cloud free during the satellite revisit; thus, relying on it for a cloud-free observation is not a solution [1]. NASA launched Terra and Aqua satellites with moderate resolution imaging spectroradiometer to study several meteorological phenomena [2]. However, studies indicate that more than 67% of global coverage and 55% of land coverage are impaired by clouds [1]. The empirical evaluations of the Brazilian Amazonian Landsat's data archive also highlight the risk of clouds in studying biophysical phenomena [3]. These undesirable cloud effects must be mitigated to increase the quality of Earth observations.

Filtering techniques evolved as the first solution for removing clouds from multispectral data. The spatial and frequency profiles of the clouds were studied to forge several homomorphic filtering methods for cloud removal from Landsat data [4]. Cloud-free reconstruction was made possible by the temporal interpolation methods that used NDVI and Fourier characteristics from multispectral input [5], [6]. However, these techniques were unreliable due to the uncertainty of clouds in temporal observation [7]. Spectro–temporal analysis employing machine learning techniques, including regression trees, independent component analysis (ICA), support vector machines, support vector regression (SVR) and multi-output SVR, outperformed traditional filtering techniques for cloud removal [8], [9], [10]. Iterative haze-optimized transformation HOT based on multivariate regression addressed the drawbacks of HOT to differentiate haze over bright land regions [11]. Compared with all previous PCA and ICA methods, a noise-adjusted principal component transform performed better [12].

Image inpainting is a popular image filling and reconstruction approach using local and global spatial dependencies. The first inpainting-based missing area reconstruction technique is a bandelet transform with multiscale geometrical grouping [13]. The spatial, spectral, and temporal relationships were included into recent inpainting-based cloud removal methods for better reconstruction [14]. Choosing the best pixels for gap-filling is the primary difficulty posed by inpainting algorithms, and this challenge can be handled with the help of auxiliary data [15].The deep learning-based cloud removal using convolutional neural networks (CNN), attention frameworks, and generative adversarial networks (GAN) enabled superior spatio–spectral feature learning from massive temporal data [16], [17], [18]. Extending the ability of attention mechanisms to traditional GANs facilitated better encoding of spatial and spectral relationships [19].

Synthetic aperture radar (SAR) imaging is carried out at longer wavelengths, and longer wavelengths penetrate through clouds and other obstacles [20]. Hence, SAR imaging is widely used for land data collection during unfavorable climatic scenarios. On the other side, clouds affect optical images, such

Manuscript received 28 December 2023; revised 8 April 2024; accepted 28 May 2024. Date of publication 7 June 2024; date of current version 1 July 2024. *(Corresponding author: Jayakrishnan Anandakrishnan.)*

Jayakrishnan Anandakrishnan and Venkatesan M Sundaram are with the Department of Computer Science and Engineering, National Institute of Technology Puducherry, Karaikal 609609, India (e-mail: jaykrizz@gmail.com; venkisakthi77@gmail.com).

Prabhavathy Paneer is with the Department of Information Technology, Vellore Institute of Technology, Vellore 632014, India (e-mail: pprabhavathy@vit.ac.in).

Digital Object Identifier 10.1109/JSTARS.2024.3411032

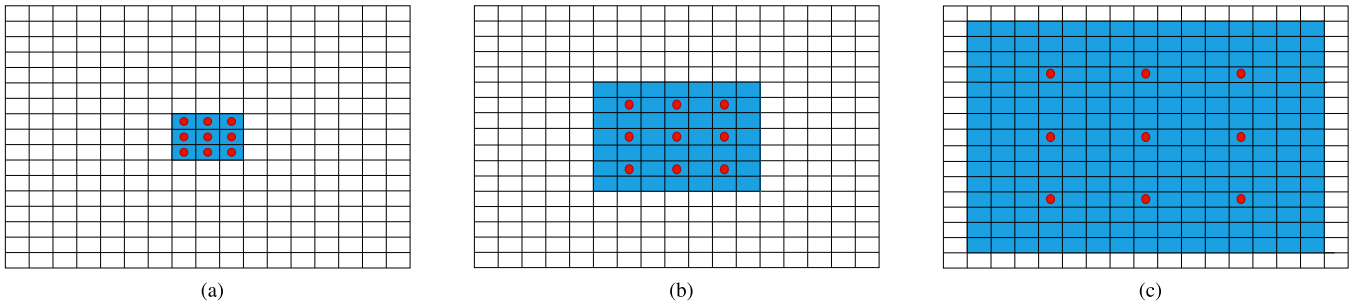


Fig. 1. $3 \times 3$ convolution carried out at different dilation rate ($p$): (a) $p = 1$, (b) $p = 2$, and (c) $p = 4$.

as hyperspectral and multispectral images, because of their wavelength regulations. The progress in multisensor data fusion techniques enhanced cloud removal paradigms by combining colocated optical and SAR data [21]. However, deep learning on these higher dimensional fused 3-D spectral data is often complex and increases model complexity [22], [23]. A spatial–temporal–spectral CNN is proposed to extract features from multisensor data, but the network failed to extract SAR physical and geometrical properties [24]. An improved conditional GAN fuses and learns coregistered SAR and temporal optical data to perform cloud removal [25]. The optical translation of SAR combined with auxiliary SAR is fed to a generative network to replace cloudy portions [26]. The first encoder–decoder architecture fusing Sentinel-1 SAR and Sentinel-2 multispectral optical data has a traditional U-Net structure with three encoders and a single decoder [27]. A convolutional–mapping–deconvolutional network performs cloud removal using low-resolution heterogeneous and SAR supplementary data [28]. A better pixelwise cloud recovery is achieved using the hybridization of convolutional long short-term memory with cGAN [29].

CNN-based cloud removal techniques are capped by their capacity to accommodate spectral and auxiliary information and fail to generalize. The existing GAN-based cloud removal performs better only on simulated data and fails for real-world scenes. Also, most of the fusion-based cloud removal techniques were underperforming due to the deficiency of proper cloud–shadow masks. The proposed Cloud Elimination from Sentinel-2 multispectral imagery using Residual Multiscale dilated Network (CERMF-Net) addresses deficiencies of exercising techniques by residual-aided multiscale learning with dilated convolutions. The notable contributions of this article are as follows.

1) Introduces a SAR-guided residual multiscale feature fusion network (CERMF-Net) with dilated convolutions for cloud elimination from Sentinel-2 data.
2) The residual blocks alleviate the vanishing gradient issue when integrating dense CNN layers for spatio–spectral feature extraction. The multiscale convolutions extract features at multiple scales and mitigate the localized feature identification problem of CNNs.
3) The receptive field of the CERMF-Net is expanded by employing dilated convolutions without any extra overhead in computational complexity. The dilations enable the network to learn global intrinsic features.
4) Proposed a cloud constrained loss function $L_{\text{ccl}}$, which prioritizes cloudy pixels in loss computation and thereby improving generalizability.

The rest of this article is organized as follows. The proposed CERMF-Net is elaborated in Section II, while Section III covers the experimental setting and outcomes. Finally, Section IV concludes this article with insightful comments and suggestions for future research.

## II. Proposed Methodology

### A. Dilated Convolution

CNN's performance is hindered by its inability to extract global features and memorize contextual information. The potential solutions, such as increasing kernel size or convolution layers, draw extra overhead in computation by increasing the number of parameters. Conversely, the dilated convolution expands the receptive field by kernel inflation. The expanded receptive field facilitates contextual feature learning. The output $y_{mn}$ of a 2-D dilated convolution with a dilation rate of $p$ on input $x$ and kernel $k$ is given in the following equation:

$$y_{mn} = f\left(\sum_{r}\sum_{i=0}^{h-1}\sum_{j=0}^{w-1} k_{ij}x_{(m-p*i)(n-p*j)} + b_{mn}\right). \quad (1)$$

The receptive field growth is exponential and is defined by the power of two at each network level ($p = 2^l$, where $l$ is the network level). That is, dilations will be carried out on two levels of $p = 4$. The receptive field of dilated convolution at various levels is obtained by the following equation, where $l_0 = 1$ by default:

$$l_i = l_{i-1} + (k-1)p. \quad (2)$$

The receptive field expansion of $3 \times 3$ dilated convolutions with various dilation rates is depicted in Fig. 1.

### B. Cloud–Shadow Map (CM) Generation

The reflectance characteristics of clouds and shadows at different wavelengths are explored to produce a CM, which enables selective reconstruction of cloud-affected regions. The method computes cloud and shadow maps separately and merges them to yield a final CM. The notations $B1$, $B2$, $B3$, $B4$, $B8$, $B10$, and $B11$ represent coastal aerosol, blue, green, red, near

infrared, short wave infrared (SWIR)-cirrus, and SWIR-2 bands of Sentinel-2.

*Cloud map generation:* The normalized difference moisture index (NDMI) and normalized difference snow index (NDSI) offer useful information in generating cloud maps [30]. The proposed cloud map generation process initially considers all pixels as cloudy ($\xi = 1$) and later updates them based on the thresholding of various spectral indices. The input to the cloud map generation is the bandwise min–max normalized Sentinel-2 reflectance data. Clouds are seen brighter in optical images. These high brightness features are thus extracted from blue, green, red, coastal aerosol, and cirrus bands of Sentinel-2 data for CM generation. In order to calculate the high reflectance-based update $\xi_{\text{HR}}$ for each pixel, the prior cloud map is sequentially replaced with the results from the following equations:

$$\begin{aligned}\xi &= \min(S_\triangledown(B2)_{[0.1:0.5]}, \text{Cloud Map})\\ \xi &= \min(S_\triangledown(B1)_{[0.1:0.3]}, \xi)\\ \xi &= \min(S_\triangledown(B10)_{[0.5:0.7]}, \xi)\\ \xi_{\text{HR}} &= \min(S_\triangledown(B2+B3+B4)_{[0.2:0.8]}, \xi).\end{aligned} \tag{3}$$

Current cloud map generation techniques rely on NDMI for most pixel estimation. NDMI-based approximations proved best for vegetation lands but failed for areas of mixed vegetation and water bodies. NDWI provides a better cloud pixel estimation for grounds with mixed vegetation and water bodies. The proposed model integrates NDWI as an auxiliary source for better evaluation of cloud pixels [31], [32]. The suggested NDWI-based cloud map update is shown in the following equation:

$$\xi_{\text{NDWI}} = \min\left(S_\triangledown\left(\frac{B3-B8}{B3+B8}\right)_{[-1:1]}, \xi_{\text{HR}}\right). \tag{4}$$

The moist pixels are identified using the NDMI index, and NDMI-based cloud pixel isolation is accomplished as follows:

$$\xi_{\text{NDMI}} = \min\left(R_\downarrow\left(\frac{B8-B11}{B8+B11}\right)_{[-0.1:0.1]}, \xi_{\text{NDWI}}\right). \tag{5}$$

Condensed water particles are often present in clouds; hence, snow pixel evaluation using NDSI is critical for cloud map generation. The NDSI-based cloud map update is given by the following equation:

$$\xi_{\text{NDSI}} = \min\left(S_\triangledown\left(\frac{B3-B11}{B3+B11}\right)_{[0.8:0.6]}, \xi_{\text{NDMI}}\right). \tag{6}$$

The rescaling operation $S_\triangledown$ performs a pixel-level remapping to the range of [L, U] and is defined by the rescaling using (7). The rescale operation carries out a pixel-level remapping between the range of [0,1], and it helps to strengthen the boundaries of the image and enables finer grained selectivity [30]

$$\xi_{S_\triangledown} = \frac{\xi - L}{U - L}. \tag{7}$$

Although the cloud map is obtained, necessary morphological operations must be performed to remove gaps in the cloud region. A structuring element of size $3 \times 3$ is used for both opening and closing operations. Apart from clouds, other highly reflective ground objects need to be neglected from the resulting cloud map. An opening operation removes such highly reflective ground objects from clouds. The resulting cloud map undergoes an opening operation to fill gaps within the cloudy regions. Finally, the cloud map is rescaled to the range of [0,1] and thresholded using a threshold value of $T_a = 0.2$. $T_a$ binarizes the resultant cloud map and helps to find the brightness component, and the final cloud map update is given as

$$\xi = \begin{cases} 1, & \text{if} \quad \xi > T_a \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

*Shadow map generation:* The shadow map generation process exploits the low brightness and low NDVI characteristics of shadow pixels. NIR, SWIR-1, coastal aerosol, and SWIR-2 bands undergo a thresholding process for extracting the shadow map [30], [33]. The average of bands B1, B8, B11, and B12 is thresholded ($T_b = 0.75$) to extract shadow pixels. Pixels with low NDVI values ($T_c = 0.8$) are assumed as dark pixels caused by the presence of water and discarded from the shadow map. The thresholds values for $T_a$, $T_b$, and $T_c$ are chosen from the experimental studies performed for Sentinel-2 data [33]. Finally, a closing operation fills the gaps in the shadow map regions and constitutes the final shadow map. The following equations provide the NDVI computation and following shadow map correction:

$$\text{NDVI} = \frac{B8-B4}{B8+B4} \tag{9}$$

$$\Phi = \begin{cases} 1, & \left(\frac{B1+B8+B11+B12}{4}\right) < T_b \text{ AND} \\ & \text{NDVI} < T_c \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

A unified CM is obtained by the pixel-wise OR operation on the individual cloud and shadow map. The process flowchart for creating the CM is displayed in Fig. 2.

### C. Cloud Constrained Loss Function

The loss function is a constraint that directs the proper fitting of the model. $L_1$ loss is a widely accepted loss function and tries to minimize the sum of all the absolute differences between the actual and predicted pixels. The $L_1$ loss function is given by the following equation, where *X* and *Y* stand for the original and restored images, respectively, and *N* denotes the total number of observations:

$$L_1 = \frac{\sum_{i=1}^{n} |X_i - Y_i|}{N}. \tag{11}$$

Temporal observations on satellite revisits constitute the reference images, and there is a high chance that the landscape will vary within the revisit period. Hence, it is critical to keep all the cloud-free background information the same. Cloud reconstruction problems only reconstruct the cloud-affected regions keeping the unaffected regions the same. The $L_1$ is computed for the overall image without giving any weightage for the cloud- and shadow-affected regions. Hence, vanilla $L_1$ loss needs to be improved for reliable reconstruction. This article

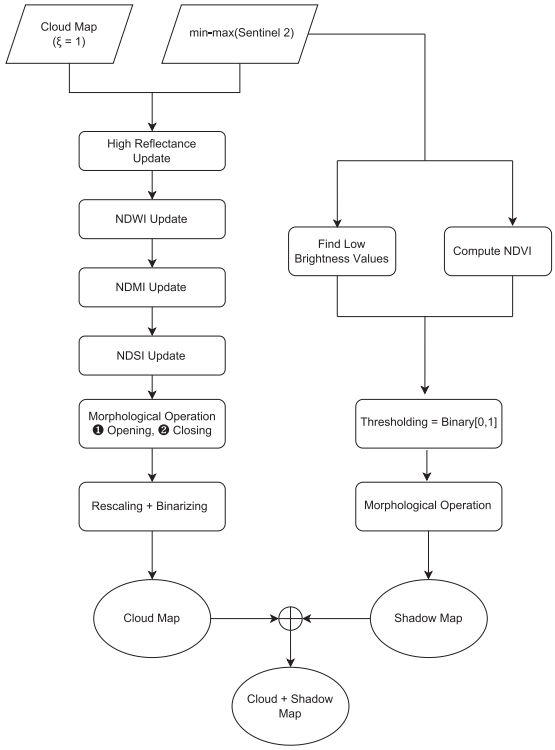


Fig. 2. Flowchart of the combined CM generation.

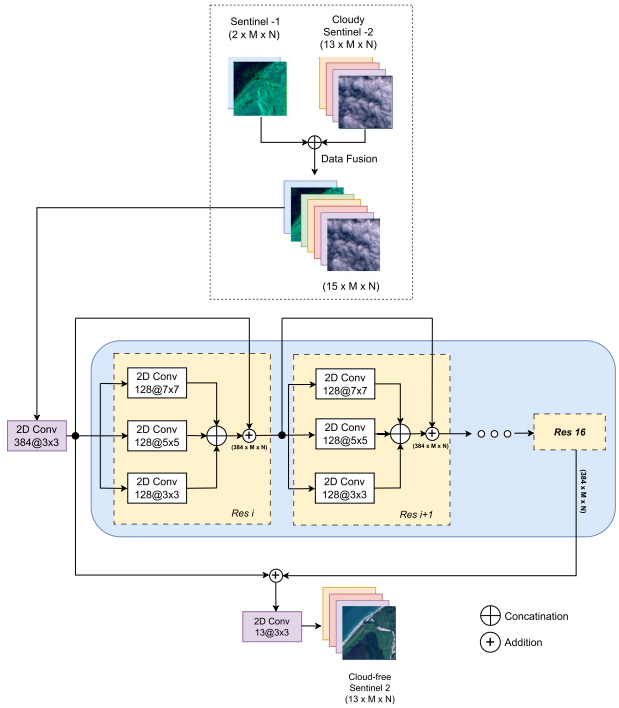


Fig. 3. Proposed CERMF-Net architecture with dilation rate $p$=2 for cloud elimination from Sentinel-2 data.

proposes a novel cloud-constrained loss function $L_{\text{ccl}}$ computed with the help of the previously generated CM. The proposed cloud-constrained loss function $L_{\text{ccl}}$ is given in the following equation:

$$L_{\text{ccl}} = \frac{\sum_{i=1}^{n} \text{CM} \cdot |X_i - Y_i|}{N} + \frac{\sum_{i=1}^{n} |X_i - Y_i|}{N}. \quad (12)$$

The $L_{\text{ccl}}$ combines cloud-constrained part and normal $L_1$. The first part of $L_{\text{ccl}}$ is constrained by the CM, which regularizes the loss computation and improves model fitness. The general $L_1$ loss section of $L_{\text{ccl}}$ is incorporated to preserve the global texture and foster smooth transitions between cloud-free and reconstructed image regions.

### D. Proposed Model

The proposed CERMF-Net employs Sentinel-1 SAR data as a complementary source of information to reconstruct cloudy regions. Sentinel-1 and Sentinel-2 datasets are represented as 3-D cubes with dimensions $B \times M \times N$, where $B$ denotes the number of bands and $M$ and $N$ represent the spatial dimensions. The channelwise concatenation of Sentinel-1 $VV$ and $VH$ bands with 13 bands of Sentinel-2 resulted in a final fused data of size $15 \times M \times N$. The fundamental principle of the CERMF-Net is extracting features at multiple scales by utilizing residual additions and dilated convolutions. The fused 15-channel input is passed to a 2-D convolution layer having 384 filters of size $3 \times 3$, producing 384 feature maps to feed succeeding residual blocks. The model is designed to have 16 residual blocks, each with three 2-D convolution branches with varying window sizes, $7 \times 7$, $5 \times 5$, and $3 \times 3$, respectively. Each convolution branch extracts 128 feature maps and concatenates them to 384 feature maps. These feature maps undergo elementwise addition with the 384 feature maps from the initial convolution. These residual addition operations are performed for every 16 residual blocks. The model employs a final long skip connection whose output is fed to a 2-D convolution layer having 13 filters of $3 \times 3$ size to produce a final cloud-free image. All the convolutions in the CERMF-Net employ rectified linear unit as an activation function. All convolution layers in the suggested model employ the same padding method to preserve the feature map size before and after convolution, with the dilation factor $p = 2$. The model is trained using the cloud-constrained loss function $L_{\text{ccl}}$ in an end-to-end manner. Fig. 3 illustrates the detailed framework of the suggested CERMF-Net.

## III. Results and Discussions

### A. Dataset

The proposed CERMF-Net is trained and tested on an all-season multiregion, multimodel dataset SEN12MS-CR [34]. The SEN12MS-CR dataset is curated from multiple regions of the globe and captured during various seasons, covering a wide variety of land. These dataset variations can ensure the dataset's validity and the model's generalizability in real-world scenarios. SEN12MS-CR is a free cloud-removal dataset available in the public domain and is carefully curated from the original SEN12MS dataset [35]. The dual-pol SAR sensor mounted on the Sentinel-1 satellite operates on interferometric wide swath mode, delivering a cloud-free $VV$ and $VH$ SAR composite. The Sentinel-2 multispectral data collected from Level-1 C top

of the atmosphere reflectance product have 13 spectral bands. The dataset has 175 nonoverlapping regions of interest (ROIs) of colocated Sentinel-1 SAR and Sentinel-2 multispectral data. The global sampling and all-season data collection concrete the credibility of the dataset, making it a benchmark dataset. Each ROI is reregistered to 10 m spatial resolution, covers a ground area of $52 \times 40$ km, and has $5200 \times 4000$ pixel resolution. The patch extraction process generates 700 patches of size $256 \times 256$ from each ROI. Finally, patches undergo boundary artefacts and distortion removal, resulting in 122 218 patch triplets. Sentinel-1 SAR, Sentinel-2 multispectral cloudy, and Sentinel-2 MS cloud-free observations recorded in the same meteorological season constitute the triplets.

### B. Experimental Setup and Parameter Setting

The experiments and model training were conducted on an NVIDIA DGX A100 GPU SERVER with 320 GB memory. The Sentinel-2 dataset is clipped to the range of [0,10 000] to exclude overbright noisy pixels. The Sentinel-1, on the other hand, undergoes value clipping within the range of $[-25.0, -32.5]$ to mitigate the presence of noise anomalies and invariant pixels. The dataset is divided into training, testing, and validation sets, with 70:15:15 proportions, and processed with min–max normalization, data augmentation, and shuffled training. The augmentation method performs shape-preserving 180° flips and 90° rotations on-the-fly using the Keras default "ImageDataGenerator" class. The input patch spatial resolution has been fixed as $256 \times 256$, where $M = N = 256$. The CERMF-Net cloud removal model implements 16 residual blocks and a final long skip connection. Each residual block has multiscale convolution branches of varying window sizes $W_1 = 7 \times 7$, $W_2 = 5 \times 5$, and $W_3 = 3 \times 3$. Besides the multiscale convolution branches and residual additions, CERMF-Net employs dilated convolutions. The dilation rate $p$ is selected from the set $1, 2, 4, 8, 16$, where $p = 1$ denotes a standard convolution. The hyperparameters, such as learning rate and batch size, have been fixed at 7e-6 and 16, respectively.

### C. Evaluation Parameters

The subjective evaluation of the reconstructed cloud-free images is often not practical [34]. Various pixelwise and spectral–structural closeness criteria are used to evaluate model performance objectively. The supervised nature of the CERMF-Net enables such objective evaluations. This section explains the evaluation criteria used to compare results in the proposed work. compared in the proposed work.

*1) Pixelwise Closeness Matrices. Mean Absolute Error (MAE):* MAE is a simple arithmetic averaging of the absolute errors between paired observations $X$ and $Y$. The observations $X$ and $Y$ should have an identical spatial and temporal resolution because MAE is a scale-variant pixelwise error matrix. In multispectral scenarios, absolute error is computed over all the spectral dimensions and averaged. A minimal MAE signifies a better restoration. Equation (13) defines the MAE between reference image $X$ and reconstructed cloud-free image $Y$. The variables $B$, $H$, and $W$ indicate the number of spectral bands, height, and width of the image, respectively. $x_{b,h,w}$ and $y_{b,h,w}$ denote pixel values in the appropriate spatial and spectral locations, respectively,

$$\text{MAE}(X, Y) = \frac{1}{B \cdot H \cdot W} \sum_{b=h=w=1}^{B,H,W} |x_{b,h,w} - y_{b,h,w}|. \quad (13)$$

*Mean Square Error (mse):* MSE is a pixel-based scale variant error metric that tries to find the sum of squares of error in prediction, whereas root-mean-squared error (RMSE) is the square root of mse. Attaining a lower mse and RMSE is a desirable criterion for reconstruction problems. The mse and RMSE between the reference image and the reconstructed cloud-free image are calculated using the following equations:

$$\text{mse}(X, Y) = \frac{1}{B \cdot H \cdot W} \sum_{b=h=w=1}^{B,H,W} (x_{b,h,w} - y_{b,h,w})^2 \quad (14)$$

$$\text{RMSE}(X, Y) = \sqrt{\text{mse}(X, Y)}. \quad (15)$$

*Peak Signal-to-Noise Ratio (PSNR):* PSNR is the ratio of the peak pixel value to the RMSE between actual and reconstructed data. Generally, a decibel-scaled logarithmic value is used to express PSNR. The PSNR value directly relates to the reconstruction quality, and a high PSNR indicate a better reconstruction. The PSNR between the reference cloud-free image and the generated cloud-free image is given in the following equation:

$$\text{PSNR}(X, Y) = 20 \cdot \log_{10} \left( \frac{1}{\text{RMSE}(X, Y)} \right). \quad (16)$$

*2) Spectral–Structural Closeness Matrices: Structural Similarity Index Measure (SSIM):* SSIM is an image-wise metric that exploits the neighborhood pixel information and captures crucial structural information from the data [36]. SSIM quantifies the structural similarity between the reference and reconstructed cloud-free images. Human visual perception relies on recognizing structural differences; hence, SSIM is a vital indicator to rank the image reconstruction quality. The value of SSIM ranges between $-1$ and 1, where 1 indicates the highest degree of similarity. The following equation reveals the SSIM between the reference image and the reconstructed cloud-free image:

$$\text{SSIM}(X, Y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}. \quad (17)$$

The mean and standard deviation of images $X$ and $Y$ are denoted by $\mu_x$, $\mu_y$, $\sigma_x^2$, and $\sigma_y^2$, respectively. $\sigma_{xy}$ indicate the covariance between $X$ and $Y$. The weak denominator problem is compromised by the constants $C1$ and $C2$, obtained from the equation $K_i L_i^2$, $i \in 0, 1$. $L$ denotes the dynamic range of the image, and $K_1$ and $K_2$ are set to 0.01 and 0.03 by default, respectively.

*Spectral Angle Mapper (SAM):* SAM measures the angle between the reference image and the reconstructed cloud-free image and is generally used for multispectral data [37]. The angle for each pixel in all available spectral bands is calculated and averaged to obtain a general spectral angle characterizing the similarity between two spectral data in terms of rotation.

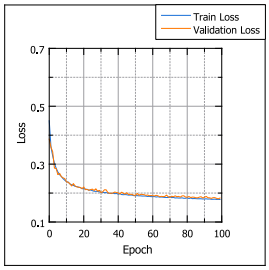


Fig. 4. Training and validation loss curves for the suggested CERMF-Net.

TABLE I
PERFORMANCE ANALYSIS OF PROPOSED MODEL VERSUS STATE-OF-THE-ART CLOUD REMOVAL TECHNIQUES USING MAE, MSE, PSNR, RMSE, SAM, AND SSIM EVALUATION METRICS

| Methods | MAE | MSE | PSNR | RMSE | SAM (°) | SSIM |
|---|---|---|---|---|---|---|
| c-GAN | **0.0441** | 0.1074 | 25.29 | 0.3278 | 14.43 | 0.7594 |
| SF-GAN | 0.0455 | 0.0886 | 24.55 | 0.2978 | 15.59 | 0.6947 |
| DSen2-CR | 0.0841 | 0.0182 | 28.70 | 0.1352 | 8.04 | 0.8750 |
| GLF-CR | 0.0821 | 0.0219 | 29.07 | 0.1481 | 7.64 | 0.8855 |
| CERMF-Net ($d$=2) | 0.0750 | **0.0162** | **32.75** | **0.1272** | **4.13** | **0.9262** |

Optimal result values are highlighted in bold case.

Hence, a smaller SAM value indicates a more closely aligned spectral reconstruction. SAM is calculated by taking the inverse cosine value of the following equation:

$$\frac{\sum_{b=h=w=1}^{B,H,W} x_{b,h,w} \cdot y_{b,h,t}}{\sqrt{\sum_{b=h=w=1}^{B,H,W} x_{b,h,w}^2 \cdot \sum_{b=h=w=1}^{B,H,W} y_{b,h,w}^2}}. \tag{18}$$

### D. Result Analysis

This study examines the performance of the proposed CERMF-Net architecture against four state-of-the-art fusion-based cloud removal methods, such as c-GAN [25], simulation-fusion GAN (SF-GAN) [26], DSen2-CR [38], and GLF-CR [39]. The traditional GAN framework is revamped to introduce c-GAN, where the generator and discriminator modules are conditioned on either SAR or temporal noncloudy observations. The two-stage SF-GAN synthesizes an optical simulation from SAR observation and later fuses it with SAR and cloudy multispectral input to generate a cloud-free composite. Although these two GAN-based techniques performed well for syntenic data, they failed for natural images, especially when applied to fully clouded areas. The DSen2-CR model has a residual neural network structure that operates on a single scale. This model performed well in reconstructing thin and thick clouds but failed in global feature extraction. Global fusion and local fusion are two crucial functions of GLF-CR. Local fusion assures the texture and details of the reconstruction, whereas global fusion ensures uniformity of the reconstruction with cloud-free areas. However, GLF-CR lacks extraction of multiscale features and sophisticated cloud–shadow mask generation procedures.

The cloud-constrained loss discovered throughout the training and validation phase of the proposed model is visualized in Fig. 4. During the early stages of training, the validation accuracy surpassed that of training. During the training, it was noted that the validation loss decreased and stabilized at a consistent level alongside the training accuracy. During epoch 97, the model demonstrated a validation loss lower than the training loss. Training has been halted at this juncture, leading to the attainment of an optimal model. Table I compares the CERMF-Net against state-of-the-art using various spectral–structural and pixelwise closeness metrics, such as MAE, mse, RMSE, PSNR, SAM, and SSIM.

The proposed CERMF-Net, with a dilation rate of 2, outperformed all the state-of-the-art methods under comparison. According to the findings, c-GAN had a low MAE score but could not obtain respectable ratings for the other evaluation metrics. All other state-of-the-art methods perform better than c-GAN, except MAE. While real-world observations with irregular clouds and noise posed challenges for GAN-based reconstruction, GAN models proved their ability to rebuild using simulated data. Compared with the newly released DSEN2-CR and GLF-CR models, the suggested cloud removal framework has a lower MAE score. The DSen2-CR model utilized the residual connection to improve the reconstruction in terms of SSIM, SAM, and PSNR. Nevertheless, the residual connections could not extract multiscale features and instead increased the number of learnable parameters. With distinct global and local feature extraction networks, the GLF-CR model outperformed DSen2-CR in terms of PSNR, SAM, and SSIM, but it also had a larger mse value. Implementing dilated convolution in the proposed CERMF-Net extracts more global features without significantly raising the number of learnable parameters. The multiscale feature extraction process combines global and local features into a single network structure. As a result, the suggested model beat all cutting-edge methods, achieving lower SAM, greater PSNR, and SSIM scores.

The subjective assessments of the suggested model against cutting-edge methods for various geographies and cloud circumstances are displayed in Fig. 5. Rows 1 and 2 of Fig. 5 represent samples of thin cloud; all the state-of-the-art algorithms were able to demonstrate fairly well cloud-free reconstruction for these samples. However, it is evident from the visualization that both c-GAN and SF-GAN struggled to maintain the structure and color accuracy. In both samples, the GLF-CR were only able to create hazy reconstruction. DSen2-CR and proposed CERMF-Net performed very well in reconstruction, but CERMF-Net showed its upper hand in retaining better color accuracy. Rows 3 and 4 represent thick clouds affected as a whole or a portion of the patch. In both cases, c-GAN and SF-GAN failed to reconstruct and produced washed-out results. GLF-CR was able to do a hazy reconstruction for a partially cloud-affected patch but entirely failed for the fully cloudy patch. In this case, DSen2-CR also performed moderately well but failed for the quality reconstruction of an entirely clouded patch. The proposed CERMF-Net showed superior reconstruction ability even for these entirely clouded patches while preserving structure and color. The results substantiate the model's capability to handle both thin and thick clouds without altering the scene's physical, texture, and color characteristics of the scene.

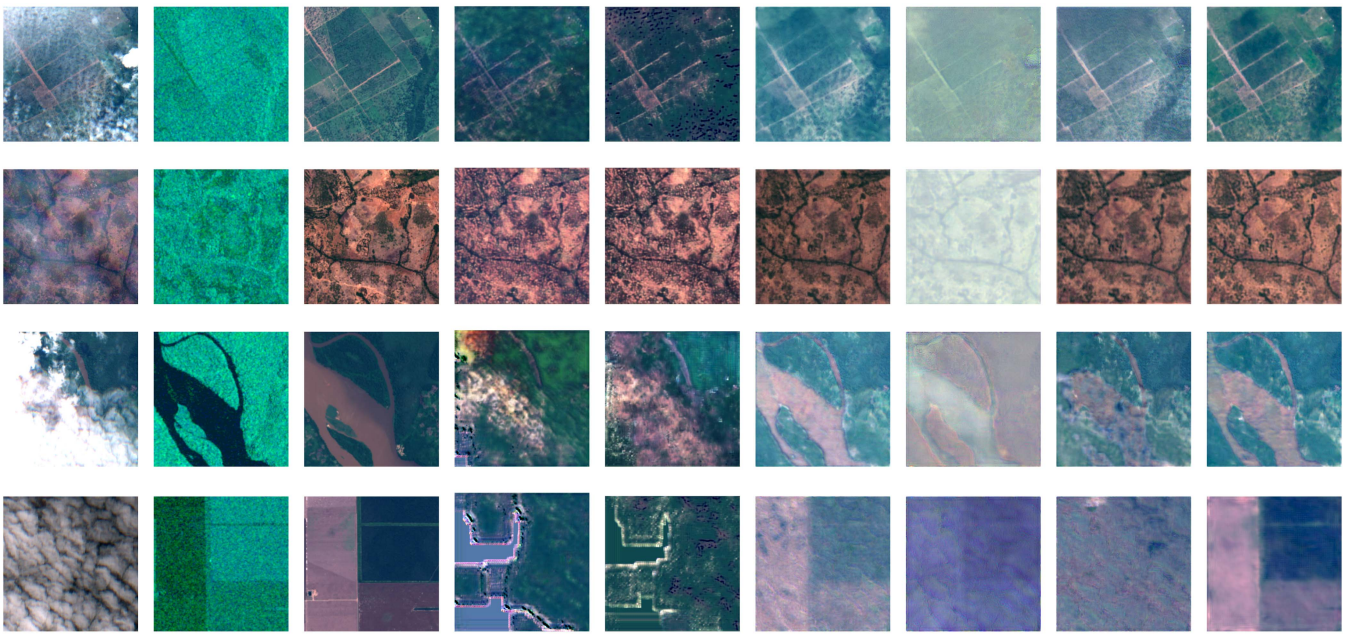


Fig. 5. Subjective comparison of proposed model with state-of-the-art techniques for different ground and cloud conditions: (a) input-cloudy, (b) input-SAR, (c) reference-cloud free, (d)–(g) reconstructed for c-GAN, SF-GAN, DSen2-CR, and GLF-CR, (h) CERMF-Net without SAR, and (i) CERMF-Net with SAR.

TABLE II
EVALUATION METRICS COMPARISON FOR THE SUGGESTED CERMF-NET WITH AND WITHOUT SAR SUPPLEMENTARY DATA

| Model Setting | MAE | MSE | PSNR | RMSE | SAM (°) | SSIM |
|---|---|---|---|---|---|---|
| Without SAR | 0.0877 | 0.0234 | 31.18 | 0.1531 | 5.32 | 0.9057 |
| With SAR | **0.0750** | **0.0162** | **32.75** | **0.1272** | **4.13** | **0.9262** |

Optimal result values are highlighted in bold case.

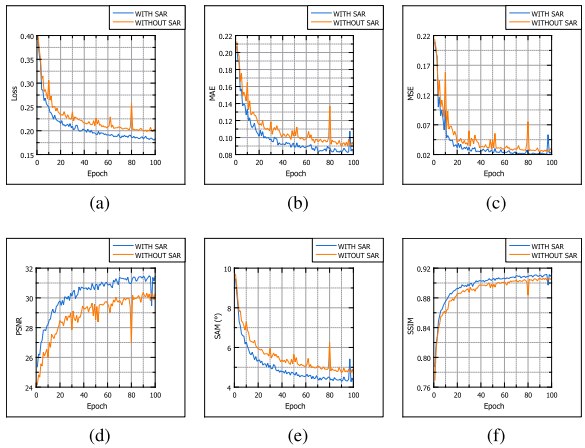


Fig. 6. Characteristic curves of (a) loss, (b) MAE, (c) mse, (d) PSNR, (e) SAM, and (f) SSIM during the validation phase for the ablation study (with and without SAR).

*1) Impact of SAR Supplementary Data on Reconstruction:* The model is trained in two different environments: one combining real cloudy data with SAR supplementary data and another excluding SAR supplementary data. This ablation study provides insight into comprehending the necessity and impact of SAR in reconstructing cloud-free imagery. The ability of SAR to penetrate through clouds facilitates the retrieval of physical information, such as boundaries, textures, and patterns, thereby enhancing the quality of reconstruction. Table II presents a quantitative comparison of evaluation metrics with and without supplementary information from SAR. Upon examination of the values presented in the table, it is evident that the reconstruction achieved through the proposed CERMF-Net in the absence of SAR is suboptimal, as evidenced by a high MAE. The elevated SAM value and diminished SSIM value suggest a reduced level of spectral resemblance compared with the network incorporating supplementary data.

Fig. 6 displays the characteristic curves of the metrics used in the ablation study, including loss, MAE, mse, PSNR, SAM, and SSIM. The graph shows that the CERMF-Net setup with supplementary SAR data worked better than the setting without SAR. The lower loss, MAE, mse, and SAM values and higher PSNR and SSIM values obtained at each epoch for the with-SAR setting substantiate the significance of SAR auxiliary data for reconstruction. The subjective assessments from subfigures (h) and (i) of Fig. 5 highlight improvements in reconstruction while incorporating SAR auxiliary data. Furthermore, compared with thin cloud removal, it is evident that thick cloud removal greatly improved with the aid of auxiliary SAR input.

*2) Selection of Dilation Factor:* Dilated convolution enables the extraction of global features by inducing an expansion in the receptive field while avoiding any additional burden on the learnable parameters. Selecting an appropriate dilation factor is crucial, and it cannot be assured that an increase in dilation rates will result in improved reconstruction. An increased dilation

TABLE III
EVALUATION METRICS COMPARISON FOR DIFFERENT DILATION RATES IN THE PROPOSED CERMF-NET

| Dilation Setting | MAE | MSE | PSNR | RMSE | SAM (°) | SSIM |
|---|---|---|---|---|---|---|
| $p$=1 | 0.0752 | 0.0167 | 32.47 | 0.1292 | 4.17 | 0.9254 |
| $p$=2 | **0.0750** | **0.0162** | **32.75** | **0.1272** | **4.13** | **0.9262** |
| $p$=4 | 0.0764 | 0.0170 | 32.65 | 0.1304 | 4.42 | 0.9172 |
| $p$=8 | 0.0795 | 0.0182 | 32.29 | 0.1348 | 4.49 | 0.9094 |
| $p$=16 | 0.0814 | 0.0210 | 31.72 | 0.1450 | 5.01 | 0.9071 |

Optimal result values are highlighted in bold case.

TABLE IV
COMPARISON OF EVALUATION METRICS FOR MODEL SETTING WITH $L_1$ AND $L_{\text{CCL}}$

| Model Setting | MAE | MSE | PSNR | RMSE | SAM (°) | SSIM |
|---|---|---|---|---|---|---|
| With $L_1$ loss | 0.0803 | 0.0271 | **32.93** | 0.1645 | 4.78 | 0.9259 |
| With $L_{\text{ccl}}$ loss | **0.0750** | **0.0162** | 32.75 | **0.1272** | **4.13** | **0.9262** |

Optimal result values are highlighted in bold case.

rate can potentially neglect the collection of local characteristics, as it may prioritize global features to a greater extent. A reduced dilation rate can capture the distinctive attributes of nearby regions, yet it is inadequate in capturing broader scale features. Therefore, conducting an empirical assessment and choosing the appropriate dilation factor are imperative. The model under consideration was tested across a range of dilation rates ($p = 1, 2, 4, 8, 16$), and the optimal value of $p = 2$ was determined based on an analysis of the evaluation metrics. As the value of $p$ exceeded 2, a decline in performance was observed. The impact of various dilation rates on assessment metrics is presented in Table III. The reconstruction deteriorates on higher levels of dilation, which can be due to a lack of local feature extraction on higher $p$.

*3) Effect of Loss Function:* The proposed model designed a new cloud-constrained loss function $L_{\text{ccl}}$, which combines the cloud-constrained part and normal $L_1$. The model is trained in two settings, one with $L_1$ loss for reconstruction, and the other with $L_{\text{ccl}}$. The setting with $L_{\text{ccl}}$ loss improves the quality of reconstruction with the aid of the cloud-constrained part and is justified by the evaluation parameters in Table IV.

## IV. CONCLUSION

Cloud cover reduces the quality of satellite observations and affects many ground applications. This article proposed a robust residual multiscale deep fusion framework with dilated convolution to remove clouds from Sentinel-2 data. The proposed CERMF-Net extracted underlying geophysical properties by fusing optical observations with auxiliary SAR data. The experimental studies revealed that thin cloud removal is possible without SAR; however, the auxiliary SAR heavily improved thick cloud reconstruction. The multiscale dilated convolutions facilitated global feature extraction with the cloud-constrained loss function obtained from the unique cloud mask. The quantitative and qualitative comparisons against diverse state-of-the-art methods ensure supreme model performance in reconstructing thick and thin clouds while retaining geometric fidelity, color, and texture. This study substantiates the necessity and impact of employing multisensor fusion techniques in the context of cloud removal tasks. One of the limiting factors in Sentinel-2 and Sentinel-1 data fusion for cloud removal is the temporal deviations in the registered data samples because of the differences in acquisition date. These temporal variations in registering samples and ground truth are neglected in this study. Future satellite programs may offer sensors simultaneous optical-SAR data registration, thus mitigating this challenge. Future study may look into combining conventional CNN with spatial attention on multitemporal fusion data to improve extraction of spatial–spectral relationship.

## REFERENCES


[1] 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.

[2] C. Justice et al., "The moderate resolution imaging spectroradiometer (MODIS): Land remote sensing for global change research," *IEEE Trans. Geosci. Remote Sens.*, vol. 36, no. 4, pp. 1228–1249, Jul. 1998.

[3] G. P. Asner, "Cloud cover in Landsat observations of the Brazilian Amazon," *Int. J. Remote Sens.*, vol. 22, no. 18, pp. 3855–3862, 2001.

[4] O. R. Mitchell, E. J. Delp, and P. L. Chen, "Filtering to remove cloud cover in satellite imagery," *IEEE Trans. Geosci. Electron.*, vol. 15, no. 3, pp. 137–141, Jul. 1977.

[5] J. Cihlar and J. Howarth, "Detection and removal of cloud contamination from AVHRR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 32, no. 3, pp. 583–589, May 1994.

[6] G. J. Roerink, M. Menenti, and W. Verhoef, "Reconstructing cloudfree NDVI composites using Fourier analysis of time series," *Int. J. Remote Sens.*, vol. 21, no. 9, pp. 1911–1917, 2000.

[7] C.-H. Lin, P.-H. Tsai, K.-H. Lai, and J.-Y. Chen, "Cloud removal from multitemporal satellite images using information cloning," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 1, pp. 232–241, Jan. 2013.

[8] E. Helmer and B. Ruefenacht, "Cloud-free satellite image mosaics with regression trees and histogram matching," *Photogrammetric Eng. Remote Sens.*, vol. 71, no. 9, pp. 1079–1089, Sep. 2005.

[9] G. Hu, X. Sun, D. Liang, and Y. Sun, "Cloud removal of remote sensing image based on multi-output support vector regression," *J. Syst. Eng. Electron.*, vol. 25, no. 6, pp. 1082–1088, 2014.

[10] Y. Shen, Y. Wang, H. Lv, and J. Qian, "Removal of thin clouds in Landsat-8 OLI data with independent component analysis," *Remote Sens.*, vol. 7, no. 9, pp. 11481–11500, 2015.

[11] S. Chen, X. Chen, J. Chen, P. Jia, X. Cao, and C. Liu, "An iterative haze optimized transformation for automatic cloud/haze detection of Landsat imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 5, pp. 2682–2694, May 2016.

[12] M. Xu, X. Jia, M. Pickering, and S. Jia, "Thin cloud removal from optical remote sensing images using the noise-adjusted principal components transform," *ISPRS J. Photogrammetry Remote Sens.*, vol. 149, pp. 215–225, 2019.

[13] A. Maalouf, P. Carre, B. Augereau, and C. Fernandez-Maloigne, "A bandelet-based inpainting technique for clouds removal from remotely sensed images," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 7, pp. 2363–2371, Jul. 2009.

[14] T.-Y. Ji, N. Yokoya, X. X. Zhu, and T.-Z. Huang, "Nonlocal tensor completion for multitemporal remotely sensed images' inpainting," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 6, pp. 3047–3061, Jun. 2018.

[15] Q. Cheng, H. Shen, L. Zhang, Q. Yuan, and C. Zeng, "Cloud removal for remotely sensed images by similar pixel replacement guided with a spatio-temporal MRF model," *ISPRS J. Photogrammetry Remote Sens.*, vol. 92, pp. 54–68, 2014.

[16] J. Li et al., "Thin cloud removal fusing full spectral and spatial features for Sentinel-2 imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 8759–8775, Oct. 2022.

[17] X. Wen, Z. Pan, Y. Hu, and J. Liu, "An effective network integrating residual learning and channel attention mechanism for thin cloud removal," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Mar. 2022, Art. no. 6507605.

[18] K. Enomoto et al., "Filmy cloud removal on satellite imagery with multispectral conditional generative adversarial nets," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 1533–1541.

[19] M. Shao, C. Wang, T. Wu, D. Meng, and J. Luo, "Context-based multiscale unified network for missing data reconstruction in remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Sep. 2020, Art. no. 8001205.
[20] H. Ghassemian, "A review of remote sensing image fusion methods," *Inf. Fusion*, vol. 32, pp. 75–89, 2016.
[21] S. C. Kulkarni and P. P. Rege, "Pixel level fusion techniques for SAR and optical images: A review," *Inf. Fusion*, vol. 59, pp. 13–29, 2020.
[22] A. Jayakrishnan, M. Venkatesan, P. Prabhavathy, and M. Alkha, "MSDF-NET: A multi-scale deep fusion network with dilated convolutions for cloud removal from Sentinel-2 imagery," in *Proc. Asia Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2023, pp. 63–70.
[23] A. Mohan and M. Venkatesan, "HybridCNN based hyperspectral image classification using multiscale spatiospectral features," *Infrared Phys. Technol.*, vol. 108, 2020, Art. no. 103326.
[24] Q. Zhang, Q. Yuan, C. Zeng, X. Li, and Y. Wei, "Missing data reconstruction in remote sensing image with a unified spatial–temporal–spectral deep convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 8, pp. 4274–4288, Aug. 2018.
[25] J. D. Bermudez, P. N. Happ, R. Q. Feitosa, and D. A. B. Oliveira, "Synthesis of multispectral optical images from SAR/optical multitemporal data using conditional generative adversarial networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 8, pp. 1220–1224, Aug. 2019.
[26] 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, 2020, Art. no. 191.
[27] R. Cresson, D. Ienco, R. Gaetano, K. Ose, and D. H. Tong Minh, "Optical image gap filling using deep convolutional autoencoder from optical and radar images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 218–221.
[28] W. Li, Y. Li, and J. C.-W. Chan, "Thick cloud removal with optical and SAR imagery via convolutional-mapping-deconvolutional network," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 4, pp. 2865–2879, Apr. 2020.
[29] A. Sebastianelli et al., "PLFM: Pixel-level merging of intermediate feature maps by disentangling and fusing spatial and temporal data for cloud removal," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Sep. 2022, Art. no. 5412216.
[30] M. Schmitt, L. H. Hughes, C. Qiu, and X. X. Zhu, "Aggregating cloud-free Sentinel-2 images with Google Earth engine," in *Proc. ISPRS Ann. Photogrammetry, Remote Sens. Spatial Inf. Sci.*, 2019, vol. IV-2/W7, pp. 145–152.
[31] X. Li et al., "Mapping water bodies under cloud cover using remotely sensed optical images and a spatiotemporal dependence model," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 103, 2021, Art. no. 102470.
[32] Y. Lin, L. He, Y. Zhang, and Z. Wu, "Cloud detection of Gaofen-2 multi-spectral imagery based on the modified radiation transmittance map," *Remote Sens.*, vol. 14, no. 17, 2022, Art. no. 4374.
[33] H. Zhai, H. Zhang, L. Zhang, and P. Li, "Cloud/shadow detection based on spectral indices for multi/hyperspectral optical remote sensing imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 144, pp. 235–253, 2018.
[34] 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.
[35] M. Schmitt, L. H. Hughes, C. Qiu, and X. X. Zhu, "SEN12MS–a curated dataset of georeferenced multi-spectral Sentinel-1/2 imagery for deep learning and data fusion," *ISPRS Ann. Photogrammetry, Remote Sens. Spatial Inf. Sci.*, vol. IV-2/W7, pp. 153–160, 2019.
[36] R. Hassen, Z. Wang, and M. M. A. Salama, "Objective quality assessment for multiexposure multifocus image fusion," *IEEE Trans. Image Process.*, vol. 24, no. 9, pp. 2712–2724, Sep. 2015.
[37] F. Kruse et al., "The spectral image processing system (SIPS)–interactive visualization and analysis of imaging spectrometer data," *Remote Sens. Environ.*, vol. 44, no. 2, pp. 145–163, 1993.
[38] 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. Photogrammetry Remote Sens.*, vol. 166, pp. 333–346, 2020.
[39] F. Xu et al., "GLF-CR: SAR-enhanced cloud removal with global–local fusion," *ISPRS J. Photogrammetry Remote Sens.*, vol. 192, pp. 268–278, 2022.

**Jayakrishnan Anandakrishnan** received the bachelor's degree in computer science and engineering from Kerala University, Kerala, India, in 2011, and the master's degree in computer science with specialization in image processing from the Cochin University of Science and Technology, Kerala, in 2013.

He is a Researcher from the National Institute of Technology Puducherry, Karaikal, India. His research interests include machine learning, remote sensing, and geospatial analytics.

**Venkatesan M Sundaram** received the bachelor's degree in computer science and engineering from the University of Madras, Chennai, India, in 1999, the master's degree in information technology, and the Ph.D. degree in computer science and engineering from VIT University, Vellore, India, in 2006 and 2014, respectively.

He is currently an Assistant Professor and the Head of the Department of Computer Science and Engineering, National Institute of Technology Puducherry, Karaikal, India. He has 23 years of teaching experience in various prestigious institutions, such as Adhiparasakthi Engg. College, Vellore Institute of Technology, and now VIT University and NIT Surathkal. In connection with his research work, he has authored or coauthored 50 articles in various reputed journals and international conference proceedings book chapters, including SCI, ESCI and Scopus. He has successfully run and executed several government-funded projects of 4 300 000 INR from ISRO, SAC, Ahmedabad, India, and DST-ICPS. His research areas include AI, machine learning, data science, data analytics, geo-spatial analysis, computational intelligence, blockchain, information retrieval, and social network analysis.

**Prabhavathy Paneer** received the M.Tech. and the Ph.D. degrees in computer science from VIT University, Vellore, India, in 2008 and 2016, respectively.

She is currently a Professor with the School of Information Technology and Engineering, VIT University, Vellore, India. She is also a part of various school activity committees. She had completed funded project from ISRO, SAC, Ahmedabad, India. She has authored or coauthored around 20 journal papers in her research field, and number of papers in international conferences. Her research interests include computational intelligence, data mining, data science, machine learning, and deep learning.

Prof. Prabhavathy is a Life Member of CSI.