SEN12MS-CR-TS: A Remote-Sensing Data Set for Multimodal Multitemporal Cloud Removal

Patrick Ebel, Yajin Xu, Michael Schmitt, *Senior Member, IEEE*, and Xiao Xiang Zhu, *Fellow, IEEE*

***Abstract*— About half of all optical observations collected via spaceborne satellites are affected by haze or clouds. Consequently, cloud coverage affects the remote-sensing practitioner's capabilities of a continuous and seamless monitoring of our planet. This work addresses the challenge of optical satellite image reconstruction and cloud removal by proposing a novel multimodal and multitemporal data set called SEN12MS-CR-TS. We propose two models highlighting the benefits and use cases of SEN12MS-CR-TS: First, a multimodal multitemporal 3-D convolution neural network that predicts a cloud-free image from a sequence of cloudy optical and radar images. Second, a sequence-to-sequence translation model that predicts a cloud-free time series from a cloud-covered time series. Both approaches are evaluated experimentally, with their respective models trained and tested on SEN12MS-CR-TS. The conducted experiments highlight the contribution of our data set to the remote-sensing community as well as the benefits of multimodal and multitemporal information to reconstruct noisy information. Our data set is available at *https://patrickTUM.github.io/cloud_removal*.**

***Index Terms*— Cloud removal, data fusion, image reconstruction, sequence-to-sequence, synthetic aperture radar (SAR)-optical, time series.**

I. Introduction

THE majority of our planet's land surface is covered by haze or clouds [1]. Such atmospheric distortions impede the capability of spaceborne optical satellites to reliably and seamlessly record noise-free data of the earth's surface. The presence of clouds is detrimental to typical remote-sensing applications, for instance, land cover classification [2], semantic segmentation [3], [4], and change detection [5], [6].

Hence, the need for cloud-free earth observation gave rise to a rapidly growing number of haze and cloud removal methods [3], [7]–[14]. Most previous methods focus on a multimodal approach [8], [13]–[15] to reconstruct cloud-covered pixels via information translated from synthetic aperture radar (SAR) or other sensors more robust to atmospheric disturbances [16], yet focus on only a single time point of observations. In comparison, recent models attempt a temporal reconstruction of cloudy observations by means of inference across time series [12], [17], [18], utilizing the circumstance that the extent of cloud coverage over a particular region is variable over time and seasons [1].

The work at hand aims to combine both preceding approaches and thus considers the challenge of cloud removal in optical satellite imagery by integrating information across time and within different modalities. For this purpose, we curate a new data set called SEN12MS-CR-TS, which contains multitemporal and multimodal satellite observations. Specifically, SEN12MS-CR-TS consists of 1-year long time series of coregistered radar Sentinel-1 (S1) as well as multispectral Sentinel-2 observations (S2) acquired in a paired manner, covering regions of interest (ROIs) from all over the world. We highlight the benefits of the proposed data set by training and testing two different models on our data set: First, a multimodal multitemporal 3-D-Convolution Neural Network that predicts a cloud-free image from a sequence of cloudy optical and radar images. Second, a sequence-to-sequence translation model that predicts a cloud-free time series from a cloud-covered time series. Both approaches are evaluated experimentally, with their respective models trained and tested on SEN12MS-CR-TS. Exemplary outcomes are highlighted in Fig. 1. The conducted experiments highlight the contribution of our curated data set to the remote-sensing community as well as the benefits of multimodal and multitemporal information to reconstruct noisy information.

A. Related Work

As the presence of clouds in optical satellite imagery poses a severe hindrance for remote-sensing applications, there has been plenty of preceding research on cloud removal methods [3], [7]–[10], [12]–[14], [20]. The focus of this overview is on data sets for cloud removal methods. Much of the

Manuscript received September 18, 2021; revised January 1, 2022; accepted January 20, 2022. Date of publication January 25, 2022; date of current version March 17, 2022. This work was supported in part by the Federal Ministry for Economic Affairs and Energy of Germany in the Project "AI4Sentinels—Deep Learning for the Enrichment of Sentinel Satellite Imagery" under Grant FKZ50EE1910. The work of Xiao Xiang Zhu was jointly supported in part by the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme under Grant ERC-2016-StG-714087, Acronym: *So2Sat*, in part by the Helmholtz Association through the Framework of Helmholtz AI under Grant ZT-I-PF-5-01—Local Unit "Munich Unit @Aeronautics, Space and Transport (MASTr)" and Helmholtz Excellent Professorship "Data Science in Earth Observation—Big Data Fusion for Urban Research" under Grant W2-W3-100, in part by the German Federal Ministry of Education and Research (BMBF) in the Framework of the International Future AI Laboratory "AI4EO—Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond" under Grant 01DD20001, and in part by the German Federal Ministry of Economics and Technology in the Framework of the "National Center of Excellence ML4Earth" under Grant 50EE2201C. *(Corresponding author: Xiao Xiang Zhu.)*

Patrick Ebel and Yajin Xu are with Data Science in Earth Observation, Technical University of Munich, 80333 Munich, Germany (e-mail: patrick.ebel@tum.de).

Michael Schmitt is with the German Aerospace Center, Remote Sensing Technology Institute, 82234 Wessling, Germany, and also with the Chair of Earth Observation, Bundeswehr University Munich, 85577 Neubiberg, Germany (e-mail: michael.schmitt@unibw.de).

Xiao Xiang Zhu is with Data Science in Earth Observation, Technical University of Munich, 80333 Munich, Germany, and also with the German Aerospace Center, Remote Sensing Technology Institute, 82234 Wessling, Germany (e-mail: xiaoxiang.zhu@dlr.de).

Digital Object Identifier 10.1109/TGRS.2022.3146246

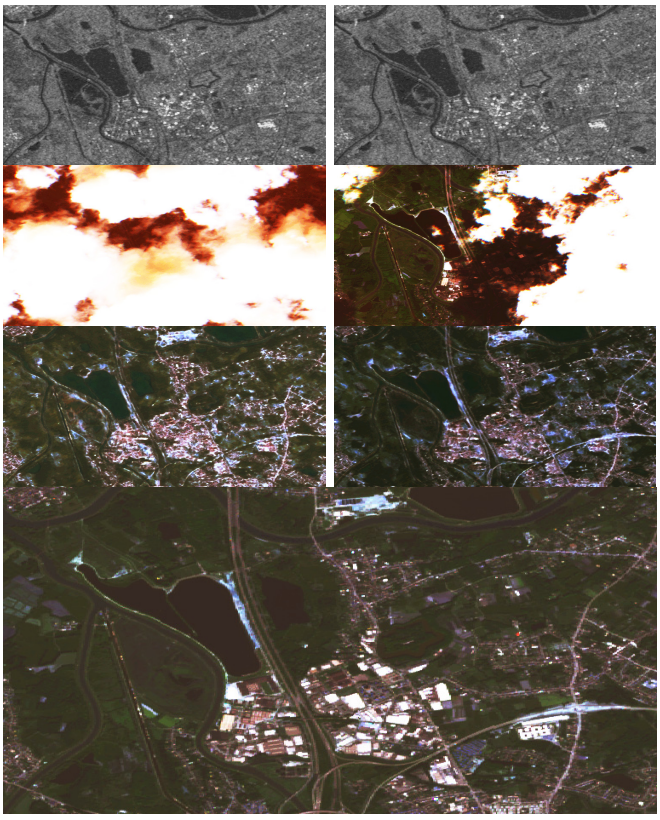


Fig. 1. Example observations and cloud-free predictions. Columns: Samples at two different time points. Rows: S1 data (in grayscale), cloudy S2 data (in RGB), predicted cloud-free $\hat{S}2$ data, and reference cloud-free S2 data of a later point in time. The results highlight that our network is able to integrate multimodal and multitemporal information to predict a clear-view sequence of multispectral observations, even in the presence of heavy cloud coverage.

early work on cloud removal considered data of simulating cloudy observations [3]. Copying cloudy pixel values from one image to another clear-view one [3] captures the spectral properties of naturally cloudy observations more faithfully than synthetic noise (e.g., Perlin noise [21]) [7], [8], [20], but neither precisely reproduce the statistics of satellite images containing natural cloud occurrences [14]. Consequently, our data set contain cloud-free as well as naturally occurring cloud-covered optical satellite recordings. The SEN12MS-CR data set [14] provides a globally distributed collection of coregistered mono-temporal Sentinel-1 as well as cloudy and cloud-free Sentinel-2 observations. Our data set is an extension of SEN12MS-CR in the sense that we collect repeated measures per ROI and therefore provide a time series of coregistered S1 and S2 observations, gathered such that matched observations of both modalities are no more than two weeks apart. In comparison to the preceding data set, ours allows integrating information not solely across different sensors, but also across different points in time distributed throughout the year. Similarly, the work of [12] allows for time-series cloud removal by providing a collection of tri-temporal RGB (NIR)-channel optical data and corresponding models. Our contribution extends this work by providing true multimodal data recorded by two distinct sensors, SAR Sentinel-1 measurements, as well as 13-band multispectral Sentinel-2 observations. Furthermore, the length of each time series is increased considerably, from 3 to 30 samples. Finally, [12] exclude observations with greater than 30% cloud coverage from their data set, which deviates from real conditions. Our approach aims to model the complete spectrum of cloud coverage, including conditions commonly encountered by remote-sensing practitioners. In sum, our work and its main contribution, a large-scale multimodal multitemporal data set for cloud removal in optical satellite imagery, build on a history of research and improve upon the current state of image reconstruction in remote sensing by providing a novel, carefully curated data set.

II. Data

This work introduces SEN12MS-CR-TS, a multimodal and multitemporal data set for training and evaluating global and all-season cloud removal methods. The data set consists of 53 globally distributed ROI, curated as detailed in Section II-A. The ROIs are over 4000×4000 px^2 each, covering about 40×40 km^2 of land such that the total surface area covered by the data set is over 80000 km^2. Of all collected ROI, 40 are defined as a training split and 13 as a hold-out split to evaluate cloud removal approaches on. For every ROI, we collect 30 coregistered and paired S1 and S2 full-scene images evenly spaced in time throughout the year of 2018. Each acquired image is inspected and quality-controlled manually. The spatial distribution of all ROI is depicted in Fig. 2 and highlights the global sampling of our data set. The empirical distribution of the cloud coverage of all optical observations (examples are shown in Fig. 3) is computed as detailed in Section II-C and the statistics are presented in Figs. 4 and 5 for the train and the test splits, respectively. The cloud-free Sentinel-2 (RGB-channel) observations of four example ROI illustrating the diversity of our data set are illustrated in Fig. 6. Importantly, the data set is curated without excluding any interval of cloud coverage such that the collected observations also reflect conditions of high cloud coverage as commonly encountered in practice [1]. The data is made available under *https://patrickTUM.github.io/cloud_removal*. It is about 2 Tb in size and compatible with the SEN12MS-CR data set [14]. That is, no train ROI of SEN12MS-CR is part of our data set's test ROI and vice versa.

A. Data Collection

All curated data are recorded via the SAR Sentinel-1 and multispectral Sentinel-2 (level 1-C top-of-atmosphere reflectance product) instruments of European Space Agency's (ESA's) Copernicus mission. The recorded observations are acquired via Google Earth Engine [22] and a custom semiautomatic processing pipeline. We randomly sample the geospatial locations of 53 ROIs from SEN12MS-CR [14]. To minimize mosaicing, observations of cells covered by a single pass are collected. The samples are referenced within the World Geodetic System 1984 (WGS84) coordinate system. For every ROI, 30 time intervals are evenly spaced throughout the year of 2018. For every time interval, a coregistered, geo-referenced, and full-scene S1 image as well as a paired full-scene S2

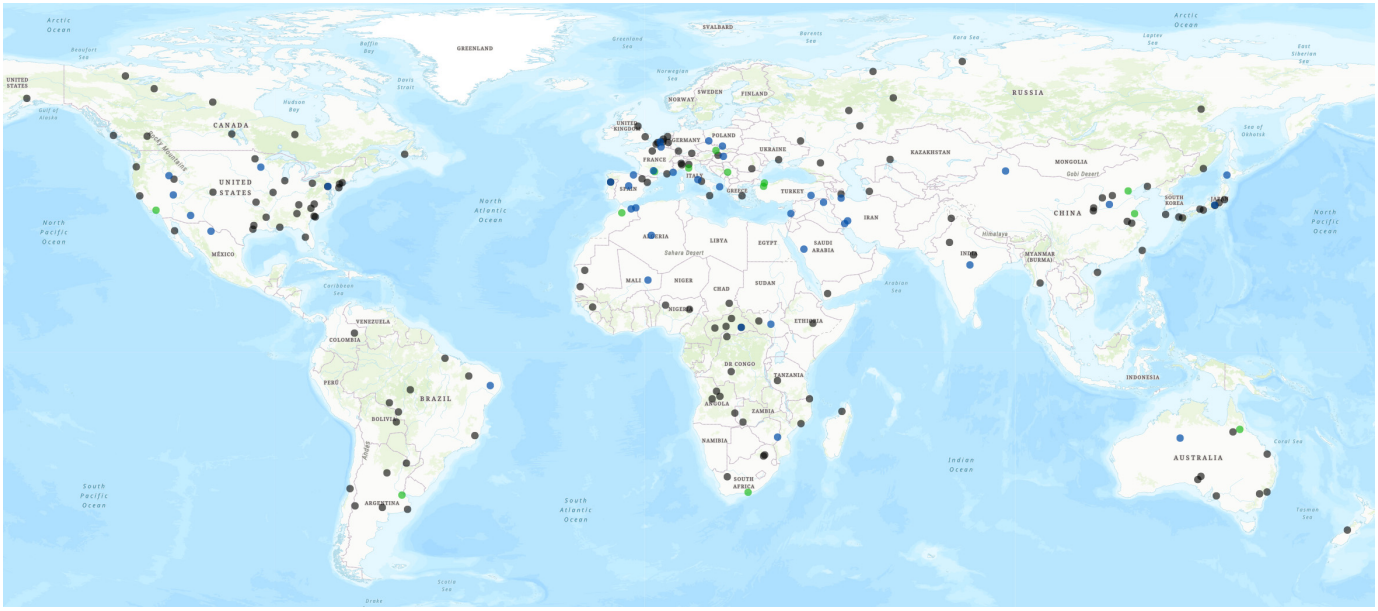


Fig. 2. Spatial distribution of the ROI constituting SEN12MS-CR-TS. Areas belonging to the training split are denoted in blue, and regions of the testing split are colored in green. The ROIs of SEN12MSCR [14], nonoverlapping and compatible with our data set, are depicted in gray. Any graphical overlap of the semitransparently plotted dots is rendered in darker tones so close-by dots can easier be discerned.

image (level 1-C) are collected. The acquisition within the same interval window is such that corresponding multimodal images are no more than two weeks apart. Further statistics regarding the pairing of observations are provided in appendix.

B. Preprocessing

To prepare the collected raw data and translate it into a format that neural networks for cloud removal can handle the following preprocessing steps are taken: Each band of every observation is upsampled to 10-m resolution (i.e., to the native resolution of Sentinel-2's bands 2, 3, 4, and 8). Every full-scene image is sliced into nonoverlapping patches of dimensions 256×256 px^2. The S1 observations are processed via the Sentinel-1 toolbox [23] (including border and thermal noise removal, radiometric calibration, and orthorectification) and decibel-transformed. An example patch-wise tuple of paired S1 and S2 data is illustrated in Fig. 3. Input patches to any ResNet model [24] are preprocessed in line with the pipeline of [13] as follows: the vertical-vertical (VV) and vertical-horizontal (VH) channels of S1 observations are value-clipped in the ranges $[-25; 0]$, $[-32.5; 0]$ and rescaled to the interval $[0; 2]$, while S2 patches are value-clipped to $[0; 10000]$ and normalized to the range $[0; 5]$. For all other networks with a different backbone architecture, preprocessing is done as follows: each patch is value-clipped and then rescaled for every pixel to take normalized values within the unit range of $[0, 1]$. The modalities S1 and S2 are value-clipped within the intervals of $[-25; 0]$ and $[0; 10000]$, respectively. This way, we follow the preprocessing protocol of the preceding work and avoid any unnecessary adjustments, for the sake of simplicity. For evaluation, the pixel values of all input patches, target images, and predictions are remapped to the unit interval $[0, 1]$, where the goodness of predictions is assessed according to the metrics stated in Section IV-A.

C. Cloud Detection and Mask Computation

In order to analyze the statistics of cloud coverage in SEN12MS-CR-TS and to model the spatio-temporal extent of clouds, we compute binary cloud masks m. For each optical image, the masks m are computed on-the-fly via the cloud detector of s2cloudless [19], which provides a binary mask of pixel-wise values in $\{0, 1\}$ that indicate cloud-free and cloud-covered pixels, respectively. The cloud mask accuracy of s2cloudless is reported to be on par with the multitemporal classifier MACCS-ATCOR joint algorithm (MAJA) [25], but the considered detector can be applied on mono-temporal satellite observations. Note that, alternatively to s2cloudless, the masks m may be computed via a dedicated neural network for cloud detection [26], [27]. However, s2cloudless has proved to be lightweight and provides sufficient performance at little extra computational cost in run time or memory, making it an appealing cloud detector to be applied on a large-scale data set such as SEN12MS-CR-TS. Example cloud detections are illustrated in Fig. 3.

III. Methods

We consider two distinctively different methods to highlight the benefits of our curated data set and the diverse tasks it allows to approach. The first method is a neural network reconstructing cloud covered pixels in time series of multimodal data to predict a single target image acquired at a cloud-free time point. The second approach introduces a neural network that performs sequence-to-sequence cloud removal, that is, it predicts a time series of cloud-free observations the same length as the cloudy input sequence.

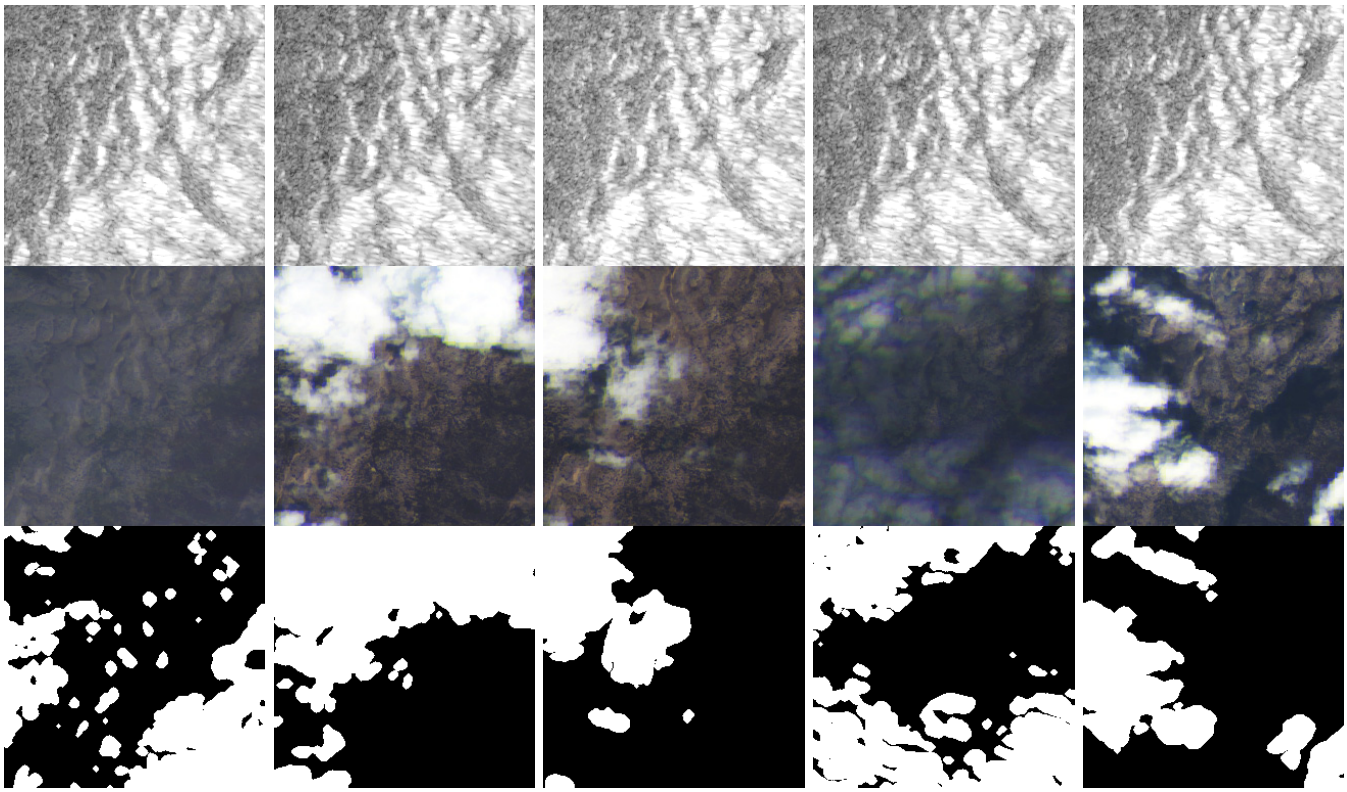


Fig. 3. Example data, preprocessed as stated in Section II-B. Rows: S1 data (in grayscale), S2 data (in RGB), and binary cloud masks (as per s2cloudless [19]). Columns: Samples of five different time points. The illustrations show that the observed region is affected by variable atmospheric disturbances and covered by a dynamic extent of clouds, changing over time. The detected cloud coverage at the individual time points is 36%, 49%, 23%, and 48%, with an average of about 39% across all illustrated samples. While some pixels are clear at least at one point in the series and may thus be reconstructed by integrating across time, whereas others are cloud-covered throughout the sequence and require spatial context or cloud-robust sensor information to be reconstructed.

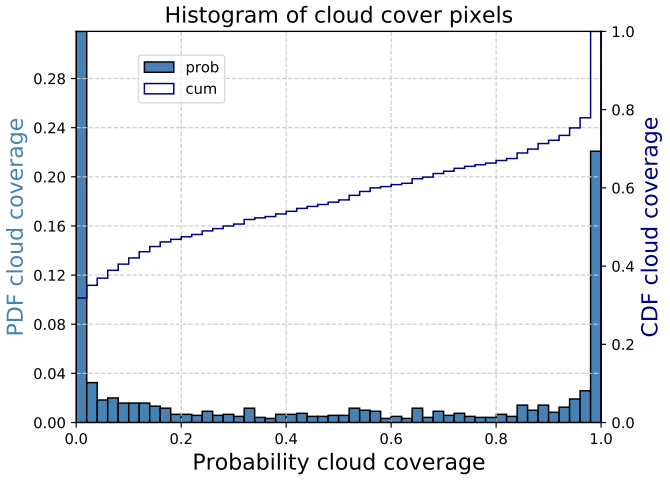


Fig. 4. Statistics of cloud coverage of SEN12MS-CR-TS train split, computed on full-scene images via the detector of [19]. On average, approximately 44% (±42%) of occlusion is observed. The empirical distribution of cloud coverage is bimodal and ranges from cloud-free views to total occlusion.

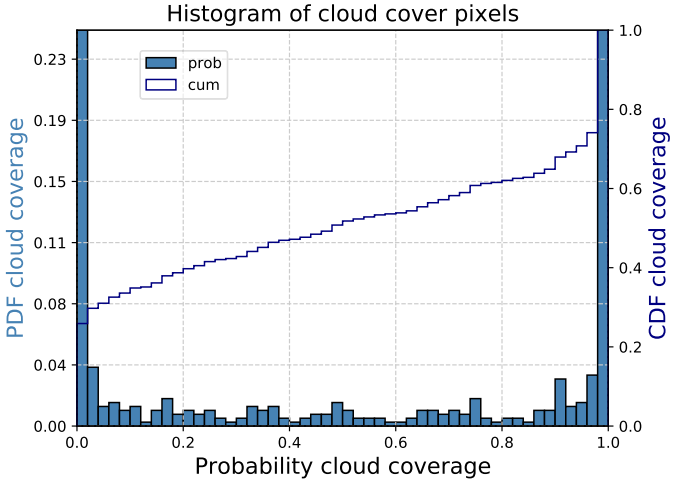


Fig. 5. Statistics of cloud coverage of SEN12MS-CR-TS test split, computed on full-scene images via the detector of [19]. On average, approximately 50% (±42%) of occlusion is observed. The empirical distribution of cloud coverage is bimodal and ranges from cloud-free views to total occlusion.

A. Multitemporal Multimodal Cloud Removal

For multitemporal multimodal cloud removal, we consider a deep neural network that builds on the generator of [12]. Our model receives a sequence of $t = 1, \ldots, n$ input tuples $(\mathrm{S1}, \mathrm{S2})_t$ and predicts a cloud-removed multispectral image $\hat{\mathrm{S2}}$. The architecture of the proposed model uses a ResNet [24] backbone, with Siamese residual branches processing the individual time points until their information gets integrated. That is, we replaced the pairwise concatenation of 2-D feature maps in [12] by stacking features in the

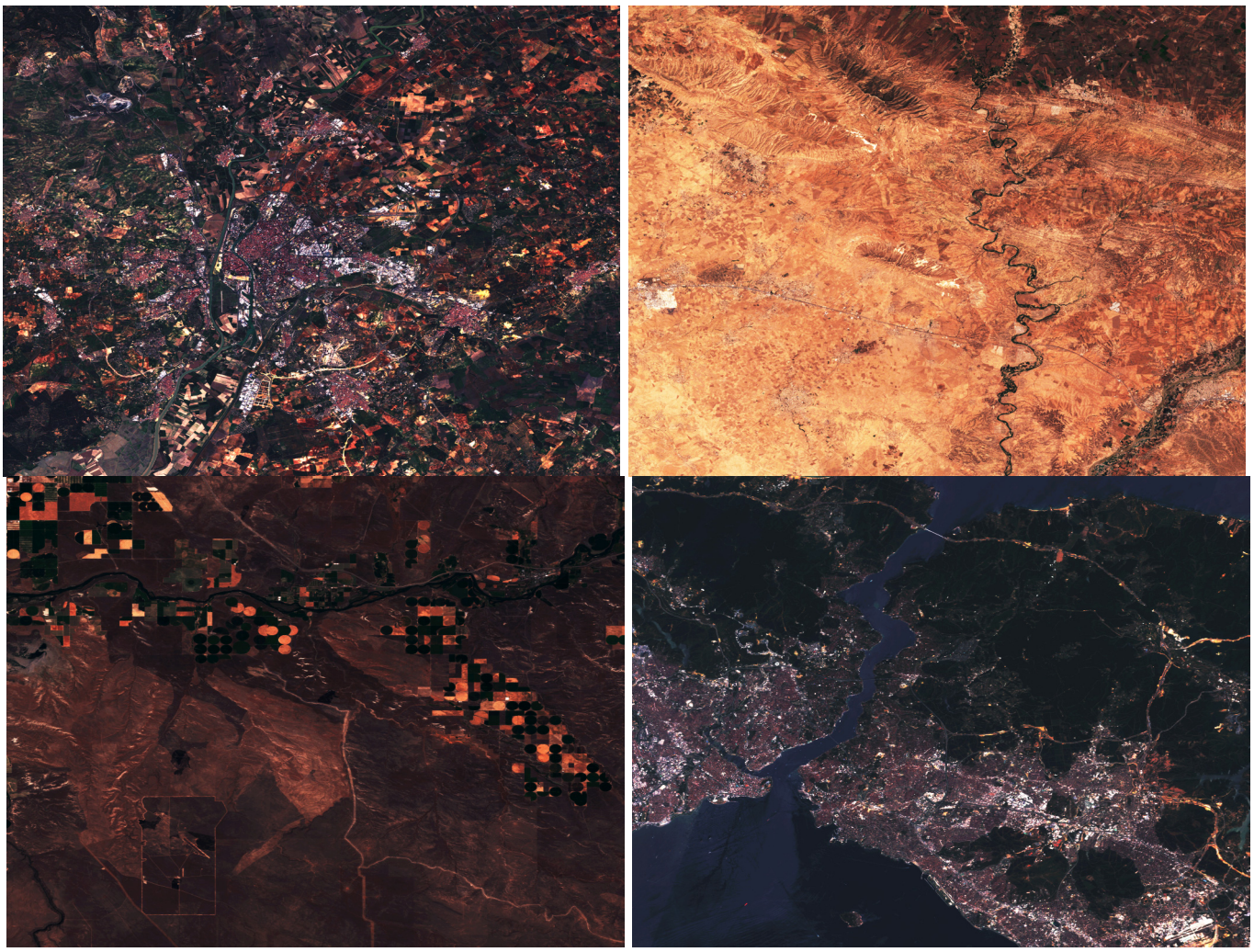


Fig. 6. Four different regions contained in SEN12MS-CR-TS, highlighting the diversity of sampled landcovers. The depicted S2 observations (RGB channels) are cloud-free samples of their respective time series. The average ROI covers about 40×40 km^2 and is split into over 700 patch samples, with each patch of size 256×256 px^2.

temporal domain, followed by 3-D convolutions. Moreover, as the first part of the generator of [12] is effectively a single time-point cloud removal subnetwork (as each time point is processed individually up to this point), we substitute this component by the established ResNet-based [24] cloud removal network of [13]. Subsequently, the feature maps are stacked in the temporal dimension and 3-D convolutions are applied to integrate information across time. The output of the network is a single cloud-free image prediction $\hat{S2}$. A schematic overview of the described architecture is shown in Fig. 7.

B. Internal Learning for Sequence-to-Sequence Cloud Removal

The sequence-to-sequence cloud removal method [28] follows the 3-D encoder–decoder architecture of [29], constituted of an encoder as well as a decoder component. Both components are arranged symmetrically in the style of U-Net [30] and linked via skip connections between paired layers. The input to the network is a sequence of multitemporal S1 samples and its output is a sequence of multitemporal cloud-removed S2 predictions. With regard to its input-to-output mapping, the proposed architecture resembles earlier SAR-to-optical translation method [31], [32]. Similar to these earlier domain translation approaches, our network learns information of the target domain (i.e., the optical imagery) via the supervision signal. Different from these approaches, the internal learning framework described below removes clouds and directly learns to denoise the target image sequence.

The architecture of the network is summarized in Fig. 8. Note that the key difference between the given model and the sequence-to-point method of Section III-A (depicted in Fig. 7) is in the output dimensions: Whereas the sequence-to-point architecture maps a sequence of n cloudy inputs to a single cloud-removed prediction, the sequence-to-sequence approach preserves the temporal information by mapping to a time series of n cloud removed outputs. Moreover, the point estimator receives tuples of S1 and S2 inputs, whereas the network of Fig. 8 is driven solely by S1 data (or Gaussian noise, as proposed in [33] and [29]). Finally, the sequence-to-point

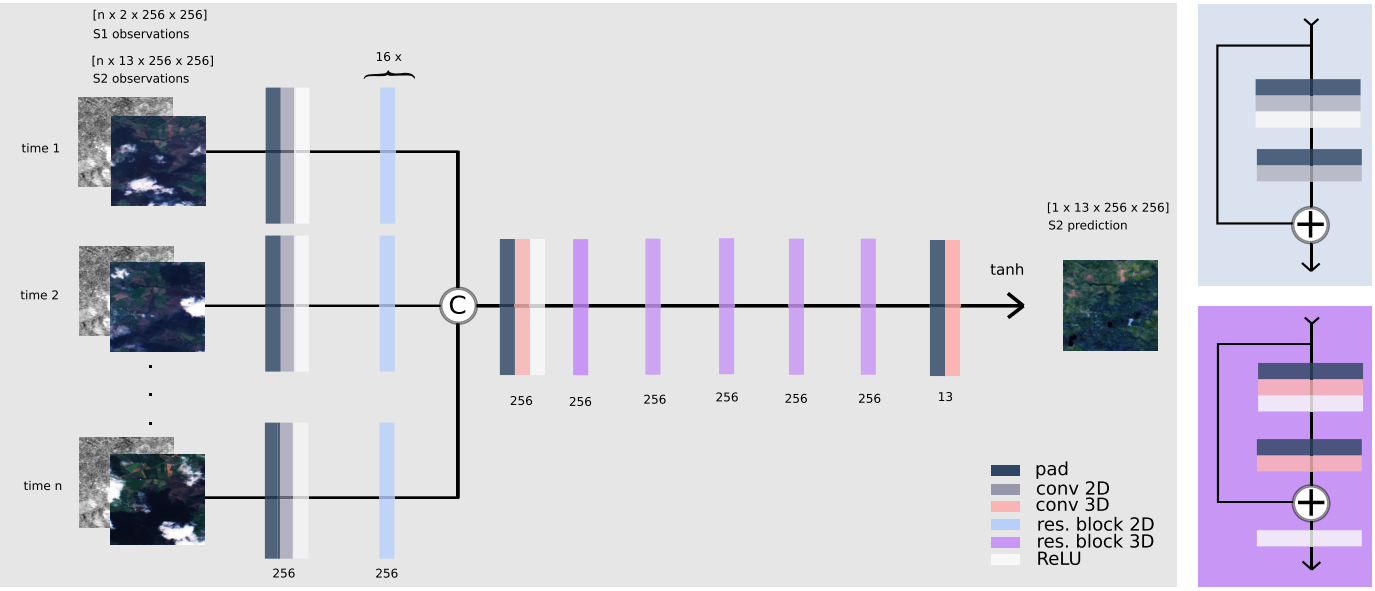


Fig. 7. Conceptual illustration of the sequence-to-point cloud removal architecture $G_{\text{seq2point}}$. The network is based on the architecture of [12] and consists of n Siamese ResNet branches [13] doing single time-point cloud removal on n individual time points. Subsequently, the feature maps are stacked in the temporal dimension and 3-D convolutions are applied to integrate information across time. The output of the network is a single cloud-free image prediction.

Algorithm 1 Internal Learning to Remove Clouds

procedure SEQ2SEQDECLOUDING($S1$, $S2$, $iterMax$)
  $G_{S1\to S2}$ = init. new NeuralNetwork()
  iterCount = 0
  while iterCount < iterMax **do**
    $\hat{S2} = G_{S1\to S2}(S1)$
    $G_{S1\to S2}.backpropagate(\mathcal{L}_{all}(S2, \hat{S2}))$
    iterCount = iterCount + 1
  Return $\hat{S2}$

network of Fig. 7 builds on the Siamese architecture of [12] with a ResNet backbone [13] plus 3-D convolutions, whereas the sequence-to-sequence approach of Fig. 8 follows a 3-D convolutional variant of U-Net [30], as proposed in [29].

The training procedure of the sequence-to-sequence network follows that of internal learning for image inpainting [29], [33], which is formalized in Algorithm 1. In this framework, for a given target sequence, a neural network is trained from scratch directly on the target sequence (without any need for additional or cloud-free training data) in order to reconstruct its noisy pixels. The observations exhibit spatio-temporal regularities and patterns (i.e., signal in the data), which is first modeled and learned by the network. The irregularities in the sequence (i.e., noise in the target data) are only internalized after, similar to a conventionally trained network overfitting to noise on training data. The internal learning approach exploits this signal–noise dichotomy and teaches a model to reconstruct cloud-covered pixels in the target sequence of S2 observations, without need for any external or cloud-free training data. In detail, a neural network is initialized and trained from scratch directly on the target sequence. At each iteration, the model receives input driving its activations (e.g., Gaussian noise or S1 recordings) and predicts a sequence $\hat{S2}$. The predictions $\hat{S2}$ are compared against the target sequence S2 (e.g., according to a cost function L_{all} as in 5) and the network learns to reproduce the cloud-free pixels. The training stops before the network overfits to internalizing the cloudy pixels.

With respect to its application and functionality, our sequence-to-sequence neural network resembles classical low-rank and sparse signal decomposition methods [34]–[37]: First, while neural networks are typically trained on a dedicated training data set separated from the test observations, numeral signal decomposition methods can be directly utilized on the data of interest. Similarly, our model can be directly applied on the test data. Second, unmixing of signals is very generic and can be applied to matrices as well as tensors. In comparison, the deep image prior approach applies to single images as well as time series [29], [33], too. Finally, the decomposition itself is into a low-rank part and a sparse component. The low-rank part denotes the data's compact representation and regularities. That is, spatial, spectral, or temporal (auto-)correlations such as the land cover mapped by a satellite. The sparse component consists of the irregular part of the data which has only a few nonzero entries, such as the appearance of clouds. In comparable terms, the internal learning technique allows our network to discover the regularities in the data and generalizing it to cloud-covered samples, before overfitting to the noise.

IV. Experiments and Results

This method details the experimental design and the corresponding results on the considered cloud removal methods as well as their ablation variants. Section IV-A specifies the measures of goodness used to assess the quality of the individual techniques' predictions. Section IV-B introduces the baselines compared against the proposed model of III-A on the sequence-to-point cloud removal task. Sections IV-C and IV-D

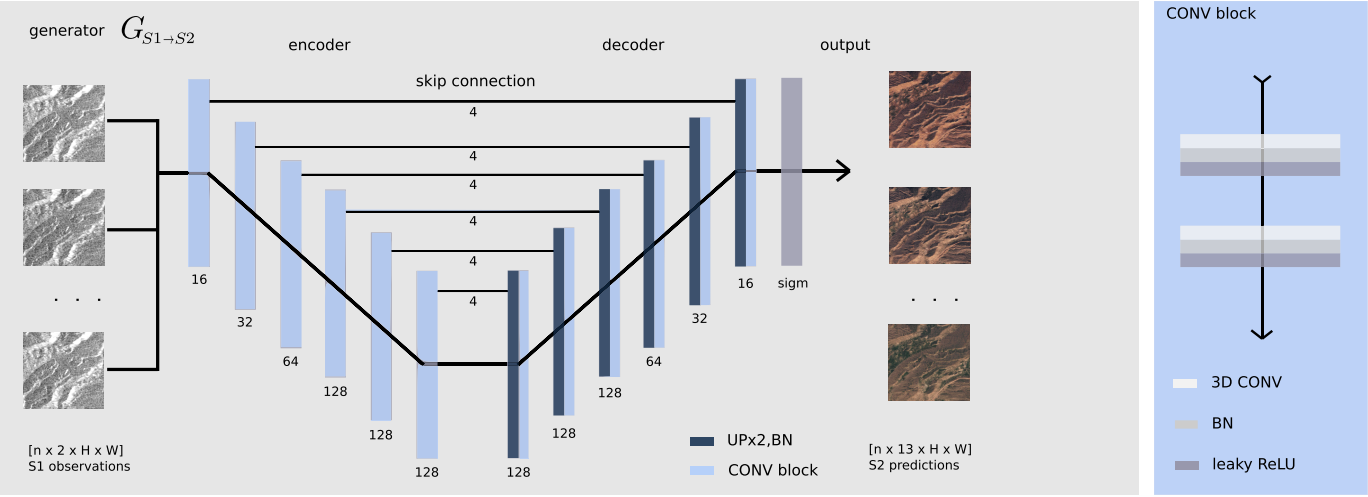


Fig. 8. Conceptual illustration of the 3-D encoder–decoder architecture G_{seq2seq} employed in the sequence-to-sequence cloud removal model [28]. The network is based on the architecture of [29] and consists of encoder and decoder parts arranged symmetrically in the style of U-Net [30], with skip connections between paired layers. Input to the network is a batch of multitemporal S1 observations. The output is a predicted batch of multitemporal multispectral S2 observations. For the ablation model considered in Section IV-D, Gaussian noise is used as an input as in [33] and [29].

detail the experiments and outcomes for the sequence-to-point and sequence-to-sequence cloud removal tasks, respectively.

A. Metrics

We evaluate the quantitative performance in terms of normalized root mean squares error (NRMSE), peak signal-to-noise ratio (PSNR), structural similarity (SSIM) [38], and Spectral Angle Mapper (SAM) [39], defined as

$$\text{NRMSE}(x, y) = \sqrt{\frac{1}{C \cdot H \cdot W} \sum_{c=h=w=1}^{C,H,W} (x_{c,h,w} - y_{c,h,w})^2}$$

$$\text{PSNR}(x, y) = 20 \cdot \log_{10}\left(\frac{1}{\text{NRMSE}(x, y)}\right)$$

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + \epsilon_1)(2\sigma_{xy} + \epsilon_2)}{(\mu_x + \mu_y + \epsilon_1)(\sigma_x + \sigma_y + \epsilon_2)}$$

$$\text{SAM}(x, y) = \cos^{-1} \times \left(\frac{\sum_{c=h=w=1}^{C,H,W} x_{c,h,w} \cdot y_{c,h,w}}{\sqrt{\sum_{c=h=w=1}^{C,H,W} x_{c,h,w}^2 \cdot \sum_{c=h=w=1}^{C,H,W} y_{c,h,w}^2}}\right)$$

with images x, y compared via their respective pixel values $x_{c,h,w}, y_{c,h,w} \in [0, 1]$, dimensions $C = 3$, $H = W = 256$, means μ_x, μ_y, standard deviations σ_x, σ_y, covariance σ_{xy} as well as constants ϵ_1, ϵ_2 to stabilize the computation. NRMSE belongs to the class of pixel-level metrics and quantifies the average discrepancy between the target and the predicted pixels in Units of the measure of interest. PSNR is evaluated on the whole image and quantifies the signal-to-noise ratio of the prediction as a reconstruction of the target image. SSIM is another image-wise measure that builds on PSNR and captures the SSIM of the prediction to the target in terms of perceived change, contrast, and luminance [38]. The SAM measure is a third image-level metric that provides the spectral angle between the bands of two multichannel images [39]. For further analysis, the pixelwise NRMSE is evaluated in three manners: 1) over all pixels of the target image (as per convention), 2) only over cloud-covered pixels (visible in neither of any input optical sample) to measure reconstruction of noisy information, and 3) only over cloud-free pixels (visible in at least one input optical patch) quantifying preservation of information. The pixel-wise masking is performed according to the cloud mask given by the detector of [19].

B. Baseline Methods

To put the performances of our proposed model and ablations into context, we consider the following baseline methods. First ("least cloudy"), taking the least-cloudy input observation and forwarding it without further modification to be compared against the cloud-free target image. This provides a measure of how hard the cloud removal task is with respect to the extent of cloud-coverage present in the data. Second ("mosaicing"), we perform a mosaicing method that averages the values of pixels across cloud-free time points, thereby integrating information across time. That is, for any pixel, if there is a single clear-view time point, then its value is copied; for multiple cloud-free samples, the mean is formed and in case no cloud-free time point exists, then a value of 0.5 is taken as a proxy. This is to avoid any extreme values, such as cloudy pixels of high intensity. The mosaicing technique provides a measure of how much information can be reconstructed across time, from multispectral optical observations exclusively. Third, ResNet refers to a residual neural network as described and trained in Sections III-A and IV-C. The architecture is based on the model of [13] and serves as a relevant baseline because parts of this model are used as Siamese residual branches within our model, as detailed in Section III-A. It provides an estimate of how well a point-to-point cloud removal model can perform as a baseline. Fourth, the baseline spatio-temporal generative adversarial network (STGAN) denotes the "Branched ResNet generator [infra-red (IR)]" architecture of [12]. It is a sequence-to-point cloud removal model, and the architecture of our own

sequence-to-point neural network closely follows its design, as detailed in Section III-A. In sum, the purpose of assessing these baselines is to analyze whether trivial solutions to the multimodal multitemporal sequence-to-point cloud removal problem exist, and how any more sophisticated deep learning approach compares against these methods and our proposed model trained on SEN12MS-CR-TS.

C. Sequence-to-Point Cloud Removal

This section details the training specifics of the sequence-to-point cloud removal architecture introduced in Section III-A. As detailed in Section III-A, up to the temporal concatenation layer, we use a version of the ResNet-based [24] cloud removal network of [13] and pretrained it on SEN12MS-CR [14] according to the training specifics of [13]. All our considered sequence-to-point cloud removal networks and ablation models share this pretrained single-temporal cloud removal network as a starting point for the sake of comparability and in order to reduce the duration of training. The networks are trained for a total of ten epochs on one tuple of patches per location for every ROI in the training split. For training, the input S2 patches are filtered to display within 0%–50% of cloud coverage. The target S2 patch is selected to be the sample showing the minimum cloud coverage over the given time series, that is, it is not necessarily temporally preceding or following the input patches. For the first 25 000 steps in the training procedure, the networks are trained with the initial ResNet Siamese components frozen, exclusively optimizing the subsequent 3-D convolution layers. After the steps with the pretrained weights frozen and once the deeper layers have been calibrated to the initial network's latent feature maps, the full network is trained end-to-end for the remainder of the process. During training, the network minimizes the loss $\mathcal{L}_{\text{all}}$

$$\mathcal{L}_{\text{all}} = \lambda_{\text{L1}}\mathcal{L}_{\text{L1}} + \lambda_{\text{perc}}\mathcal{L}_{\text{perc}} \tag{1}$$

$$\mathcal{L}_{\text{L1}} = ||\text{S2} - \hat{\text{S2}}||_1 \tag{2}$$

$$\mathcal{L}_{\text{perc}} = ||\text{VGG16(S2)}, \text{VGG16}(\hat{\text{S2}})||_2 \tag{3}$$

with $\lambda_{\text{L1}} = 100$ according to [12] and $\lambda_{\text{perc}} = 1$ as hyperparameters weighting the individual pixel-wise loss $\mathcal{L}_{\text{L1}}$ and the perceptual loss $\mathcal{L}_{\text{perc}}$. The perceptual loss is computed by means of an auxiliary Visual Geometry Group 16 (VGG16) network [40] resulting in sharper image reconstructions [41]. In comparison to other VGG16 pretrained on classical computer vision data sets such as ImageNet [42] and thus limited to RGB channel data, we pretrained a VGG16 for landcover classification on the SEN12MS data set [43] according to the training protocol of [2]. The proposed sequence-to-point cloud removal network and its ablation variants are optimized via Adaptive Moment Estimation (ADAM) [44], with a learning rate of 0.0002 and momentum parameters $[0.5, 0.999]$ as in [12]. A batch size of one tuple of samples per iteration is used for training.

To evaluate performances on the test split, samples containing S2 observations from the complete range of cloud coverage (between 0 and 100%) are considered for input. Table I compared the results of our proposed model with the baselines detailed in Section IV-B. The results show that the proposed network outperforms the baselines in the majority of metrics, except for PSNR (where mosaicing comes first) and the NRMSE (clear) preservation metric (where the "least cloudy" approach performs best). This demonstrates that a deep neural network approach can typically outperform trivial solutions to the multimodal multitemporal cloud removal problem. Exemplary outcomes for the considered baselines on four different samples from the test split are presented in Fig. 9. The considered cases are cloud-free, partly cloudy, cloud-covered with no visibility except for a single time point, and cloud-coverage with no visibility at any time point. The results show that the considered models typically outperform the simple heuristics. One exceptional case is least cloudy in the absence of clouds, which manages to accomplish a faithful prediction in such settings. Moreover, the illustrations underline that multitemporal and multimodal data may benefit image reconstruction: While most methods perform well in the cloud-free or partly cloudy cases, multisource integration is needed if individual time points contain dense cloud coverage over wide areas. When all input data is covered by thick clouds, then this poses a severe challenge for all approaches considered. To analyze the benefits of including S1 SAR data, we perform an ablation study and compare a multisensor model against one only utilizing multispectral S2 input. Table II compared the results of the multimodal model with an ablation version not using S1 SAR data. The comparison illustrates the benefits of including SAR data when reconstructing cloud-covered pixels. Next, we conduct an ablation experiment to assess the additional benefits of utilizing the introduced perceptual loss. Table III compared the results of our proposed model with an ablation version not using the perceptual loss (i.e., setting $\lambda_{\text{perc}} = 0$ in eq 1). The outcomes imply that the usage of a perceptual loss results in cloud-removed predictions of a higher quality. Finally, we consider the extension of the proposed model into networks integrating four and five time points of input information. Table IV compared the performance of our model as a function of input time points ($n = 3, 4, 5$). The results indicate that considering longer time series may provide further improvements in terms of reconstructing cloud-covered information. In a final experiment on sequence-to-point cloud removal, Table V reports the performance of our proposed model ($n = 3$, with S1 and perceptual loss) as a function of cloud coverage. That is, for a given interval of cloud

TABLE I

QUANTITATIVE EVALUATION OF THE PROPOSED SEQUENCE-TO-POINT MODEL WITH BASELINE APPROACHES IN TERMS OF NORMALIZED ROOT MEAN SQUARED ERROR (NRSME), PSNR, SSIM [38], AND THE SAM [39] METRIC. OUR MODEL PERFORMS BEST IN THE MAJORITY OF METRICS, DEMONSTRATING THAT A DEEP NEURAL NETWORK APPROACH YIELDS ADDITIONAL BENEFITS OVER TRIVIAL SOLUTIONS TO THE MULTIMODAL MULTITEMPORAL CLOUD REMOVAL PROBLEM

model	NRMSE (all)	NRMSE (cloudy)	NRMSE (clear)	PSNR	SSIM	SAM
least cloudy	0.079	0.082	**0.031**	—	0.815	0.213
mosaicing	0.062	0.064	0.036	**31.68**	0.811	0.250
ResNet	0.060	0.062	0.040	26.04	0.810	0.212
STGAN	0.057	0.059	0.050	25.42	0.818	0.219
ours (n=3)	**0.051**	**0.052**	0.040	26.68	**0.836**	**0.186**

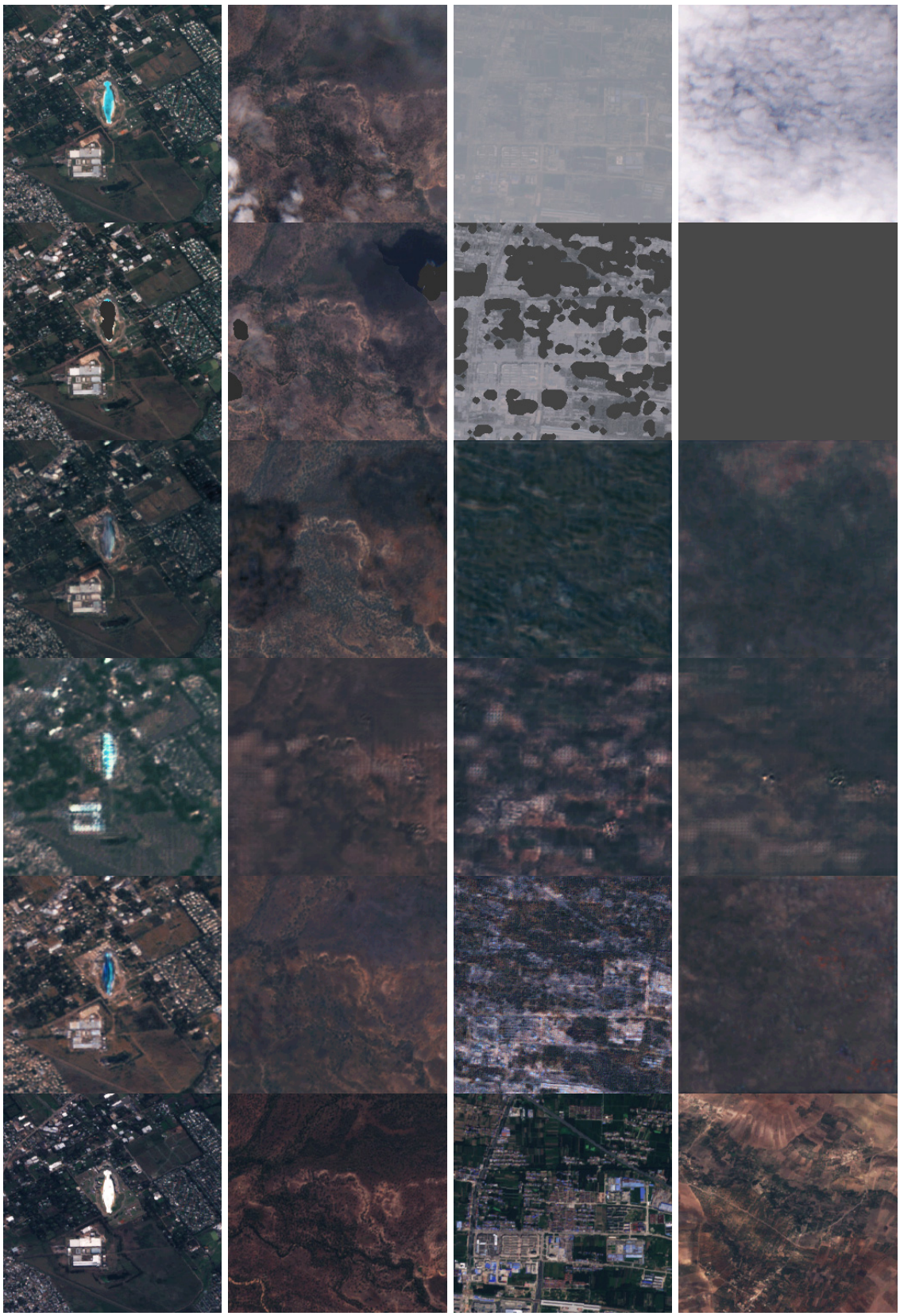


Fig. 9. Exemplary predictions and cloud-free target images for all baselines reported in Table I. Columns: Four different samples from the test split. The considered cases are cloud-free, partly cloudy, cloud-covered with no visibility except for a single time point, and cloud-covered with no visibility in any time point. Rows: Predictions of least cloudy, mosaicing, ResNet, STGAN, ours (n=3), as well as the cloud-free reference image. The results show that the considered models outperform the simple heuristics. Moreover, the illustrations underline that multitemporal and multimodal data may benefit image reconstruction.

TABLE II

COMPARISON OF THE PROPOSED SEQUENCE-TO-POINT MODEL INCLUDING SAR OBSERVATIONS VERSUS AN ABLATION VERSION WITHOUT SAR OBSERVATIONS IN TERMS OF NRSME, PSNR, SSIM [38], AND THE SAM [39] METRIC. THE COMPARISON ILLUSTRATES THE BENEFITS OF INCLUDING SAR DATA WHEN RECONSTRUCTING CLOUD-COVERED PIXELS

model	NRMSE (all)	NRMSE (cloudy)	NRMSE (clear)	PSNR	SSIM	SAM
ours (no S1)	0.054	0.057	0.054	25.35	0.832	0.194
ours (with S1)	**0.051**	**0.052**	**0.040**	**26.68**	**0.836**	**0.186**

TABLE III

COMPARISON OF THE PROPOSED SEQUENCE-TO-POINT MODEL INCLUDING PERCEPTUAL LOSS VERSUS AN ABLATION VERSION WITHOUT PERCEPTUAL LOSS IN TERMS OF NRSME, PSNR, SSIM [38], AND THE SAM [39] METRIC. THE OUTCOMES IMPLY THAT THE USAGE OF A PERCEPTUAL LOSS DURING TRAINING RESULTS IN CLOUD-REMOVED PREDICTIONS OF A HIGHER QUALITY AT TEST TIME

model	NRMSE (all)	NRMSE (cloudy)	NRMSE (clear)	PSNR	SSIM	SAM
ours (no percept.)	0.052	0.053	**0.039**	26.66	0.835	**0.180**
ours (with percept.)	**0.051**	**0.052**	0.040	**26.68**	**0.836**	0.186

TABLE IV

QUANTITATIVE EVALUATION OF THE PROPOSED SEQUENCE-TO-SEQUENCE MODEL WITH VARYING NUMBERS OF TIME POINTS ($n = 3, 4, 5$) IN TERMS OF NRSME, PSNR, SSIM [38], AND THE SAM [39] METRIC. OUR MULTITEMPORAL NETWORK WITH SAR GUIDANCE OUTPERFORMS THE MULTITEMPORAL ABLATION MODEL WITHOUT PRIOR SAR INFORMATION

model	NRMSE (all)	NRMSE (cloudy)	NRMSE (clear)	PSNR	SSIM	SAM
ours (n=3)	0.051	0.052	0.040	26.68	0.836	0.186
ours (n=4)	0.049	0.050	0.041	**27.10**	0.845	**0.172**
ours (n=5)	**0.048**	**0.048**	**0.032**	27.07	**0.846**	0.178

TABLE V

PERFORMANCE OF OUR SEQUENCE-TO-POINT CLOUD REMOVAL METHOD ($n = 3$, WITH S1 & WITH PERCEPTUAL LOSS) AS A FUNCTION OF CLOUD COVERAGE. FOR A GIVEN INTERVAL, ALL $n = 3$ INPUT IMAGES ARE SAMPLED TO CONTAIN A CORRESPONDING EXTENT OF CLOUDS. THE OUTCOMES SHOW THAT IMAGE RECONSTRUCTION PERFORMANCE IS HIGHLY DEPENDENT ON THE PERCENTAGE OF CLOUD COVERAGE. WHILE PERFORMANCE DECREASE IS NOT STRICTLY MONOTONOUS WITH AN INCREASE IN CLOUD COVERAGE, A STRONG ASSOCIATION PERSISTS

% cloud coverage	NRMSE (all)	NRMSE (cloudy)	NRMSE (clear)	PSNR	SSIM	SAM
0-10 %	**0.041**	**0.046**	**0.041**	**28.59**	**0.870**	**0.143**
10-20 %	0.044	**0.046**	0.043	27.69	0.848	0.166
20-30 %	0.046	0.047	0.044	27.25	0.841	0.169
30-40 %	0.048	0.050	0.045	26.77	0.830	0.169
40-50 %	0.047	0.048	0.045	26.86	0.830	0.167
50-60 %	0.049	0.494	0.048	26.55	0.825	0.185
60-70 %	0.052	0.052	0.043	26.10	0.817	0.184
70-80 %	0.049	0.050	0.044	26.59	0.816	0.179
80-90 %	0.050	0.050	0.044	26.54	0.820	0.175
90-100 %	0.063	0.063	—	24.79	0.786	0.222

coverage, all $n = 3$ input images are sampled to contain a corresponding extent of clouds. The outcomes show that image reconstruction performance is highly dependent on the percentage of cloud coverage. While performance decrease is not strictly monotonous with an increase in cloud coverage, a strong association persists.

TABLE VI

QUANTITATIVE EVALUATION OF BASELINE METHODS AND THE PROPOSED SEQUENCE-TO-SEQUENCE MODEL IN TERMS OF ROOT MEAN SQUARED ERROR (RSME), PSNR, SSIM, AND THE SAM [39] METRIC. OUR MULTITEMPORAL NETWORK WITH SAR GUIDANCE OUTPERFORMS THE CONSIDERED BASELINES AS WELL AS THE MULTITEMPORAL ABLATION MODEL WITHOUT PRIOR SAR INFORMATION

model	NRMSE (all)	PSNR	SSIM	SAM
RPCP [45]	0.403	7.911	0.264	30.567
NMFISL [46]	0.312	10.262	0.450	29.285
PNMF [47]	0.317	10.135	0.432	29.801
MNMF [48]	0.361	8.945	0.361	28.685
OSTD [49]	0.303	10.853	0.402	35.454
seq2seq (no S1)	0.298	11.434	0.494	28.127
seq2seq (with S1)	**0.274**	**11.590**	**0.512**	**27.733**

D. Sequence-to-Sequence Cloud Removal

A key characteristic of training the sequence-to-sequence cloud removal model described in Section III-B is the model being trained directly on the time series of images one aims to removes clouds from, without the use of any external training data as in [33] and [29]. More specifically, the training procedure teaches the network to replicate cloud-free pixels and inpaint cloud-covered ones in the target sequence S2 according to the cost function $\mathcal{L}_{\text{all}}$ formulated in [29] as

$$\mathcal{L}_{\text{all}} = \lambda_{L2}\mathcal{L}_{\text{L2}} + \lambda_{\text{perc}}\mathcal{L}_{\text{perc}} \tag{4}$$

$$\mathcal{L}_{\text{L2}} = ||\text{S2} \cdot (1-m), \quad \hat{\text{S2}} \cdot (1-m)||_2 \tag{5}$$

$$\mathcal{L}_{\text{perc}} = ||\text{VGG16(S2)} \cdot (1-m), \text{VGG16}(\hat{\text{S2}}) \cdot (1-m)||_2 \tag{6}$$

where $\lambda_{\text{L2}} = 1$ and $\lambda_{\text{perc}} = 0.01$ refer to hyperparameters that linearly combine the terms constituting $\mathcal{L}_{\text{all}}$. $\mathcal{L}_2$ is a pixel-wise reconstruction loss evaluated over the cloud-free pixels via an auxiliary VGG16 network [40] as explained before. The pseudo-code formalizing the intrinsic learning procedure is given in Algorithm 1 described in Section III-B and further justifications are stated in the original work of [33]. For a given target sequence, the network is trained for 20 passes with batches of $n = 5$ samples consisting of temporally adjacent images, for 100 iterations per pass. The network is optimized via ADAM [44] with a learning rate of 0.01 and the hyperparameters of Algorithm 1 set as stated in [29].

To quantitatively evaluate the considered model on SEN12MS-CR-TS, we propose the following protocol for a sequence-to-sequence cloud removal task: For a given target sequence, the least cloud-covered S2 observation is identified and denoted as a target image S2_t. The most cloudy S2 sample is observed and denoted as a source image S2_s. The cloud-covered pixels of S2_s according to a cloud mask m are alpha-blended with the cloud-free pixels of $S2_t$ similar to the approach of [3]. Finally, the cloud-removed prediction $\hat{\text{S2}}_t$ is then compared against the originally cloud-free S2_t in order to get a measure of goodness of cloud removal.

Table VI shows the results of the proposed network on the sequence-to-sequence cloud removal task following the

Fig. 10. Illustration of baseline methods for the sequence-to-sequence cloud removal task. The presented results show a cloudy image to be declouded, as well as the predictions via Riemannian Robust Principal Component Pursuit (RPCP) [45], Nonnegative Matrix Factorization Incremental Subspace Learning (NMFISL) [46], Probabilistic Nonnegative Matrix Factorization (PNMF) [47], Manhattan Nonnegative Matrix Factorization (MNMF) [48], and Online Stochastic Tensor Decomposition (OSTD) [49]. The results indicate that the presence of large and dense clouds poses a severe challenge for the considered methods. Most baselines decloud the image except for some residual artifacts, and some techniques display discolorization. For comparison with ours (no S1), ours (with S1), and the cloud-free target image, see Fig. 11.

aforementioned protocol. Furthermore, the considered model is compared against an ablation model, conditioned on random Gaussian noise as in [33] and [29] in place of the meaningful S1 input observations. Example outcomes of sequence-to-sequence cloud removal on a given ROI are depicted in Figs. 1 and 10. Furthermore, Fig. 11 provides a qualitative comparison between the predictions conditioned on SAR versus no prior information, underlining the benefits of multimodal information. The results highlight that the internal learning approach can learn to reconstruct cloud-covered pixels on a very limited amount of data. Furthermore, the results demonstrate that including SAR data results in performance benefits over the single-sensor baseline.

V. Discussion

The main contribution of this work is in curating and providing SEN12MS-CR-TS, a multimodal multitemporal data set for cloud removal in optical satellite imagery. Our large-scale data set covers a heterogeneous set of ROIs sampled from all over earth, acquired in different seasons throughout the year. Given that the contained observations cover clear-view, filmy, as well as nontransparent dense clouds, the objective of reconstructing cloud-covered information poses a challenging task for the considered methods and future approaches. For the sake of demonstrating the usefulness of the presented data set, we propose a sequence-to-point as well sequence-to-sequence cloud removal network. The considered methods are evaluated in terms of pixel-wise and image-wise metrics. We provide evidence that taking time-series information into account is facilitating the reconstruction of cloudy pixels and that including multisensor measurements does further improve the goodness of the cloud-removed predictions, justifying the design of SEN12MS-CR-TS to include multitemporal and multimodal data. The major difference to the preceding mono-temporal SEN12MS-CR data set [15] for cloud removal is that SEN12MS-CR-TS features a time series of 30 samples per ROI. This allows for developing methods that integrate information across time to more faithfully reconstruct cloud-obscured measurements. The sensitivity to temporal information may be particularly valuable for future research investigating the benefits of cloud removal to time-sensitive applications, such as change detection. On the other side, there is a tradeoff in terms of size, and while SEN12MS-CR-TS is more than twice as large as its mono-temporal precursor, the latter contains about two times as many ROIs sampled over all continents. However, both data sets are fully compatible, meaning that holdout ROIs of one belong to the test split of the other data set and vice versa. As there is no geo-spatial overlap across splits between both data sets, they can be combined for training or validation purposes. Finally, the two data sets exhibit a comparable extent of cloud coverage—about 50% and 48%, respectively, both covering the full spectrum from semitransparent haze to thick and dense clouds. A discrepancy between both data sets is in SEN12MS-CR having between 25% and 50% overlap between neighboring patches (following the design of [43]), whereas SEN12MS-CR-TS has no intersection between adjacent samples. SEN12MS-CR contains 122 218 patch triplets of S1, cloudy S2, and cloud-free S2 data, whereas SEN12MS-CR-TS consists of 30 time samples for each of the 15 578 patch-wise observations, for every S1 and S2 measurement. Due to the differences in preprocessing the two data sets are not coregistered patch-wise but, importantly, they share a common definition of ROIs as well as train and test splits. This way, they are compatible with one another such that SEN12MS-CR-TS can be utilized for time-series cloud removal, while SEN12MS-CR can provide further geospatial coverage of additional ROIs on individual time points. Thanks to the different designs of both data sets, they may prove beneficial facilitating a variety of downstream tasks, such as semantic segmentation [43], scene classification [2], or change detection [5], even in the presence of clouds.

Beyond the design of our novel data set, additional contributions of this work are in introducing the internal learning

Fig. 11. Illustrations on the effect of prior guidance via SAR information. Columns: SAR input to the SAR-conditioned model, cloud-free prediction of the model conditioned on Gaussian noise, cloud-free prediction of the model conditioned on SAR information, and cloud-free observation as a reference image. The structural information provided by the SAR input provides a strong prior to the model, guiding it toward learning to remove clouds in the cloudy input time series.

approach to cloud removal in optical satellite data, as well as demonstrating that SAR-to-optical cloud removal performs better than the original noise-to-optical translation framework. While our data set aims to provide a global distribution of samples, we think that the internal learning approach to cloud removal may be of particular interest for remote-sensing practitioners focusing on a single a spatially confined ROI, as no further external data is necessary.

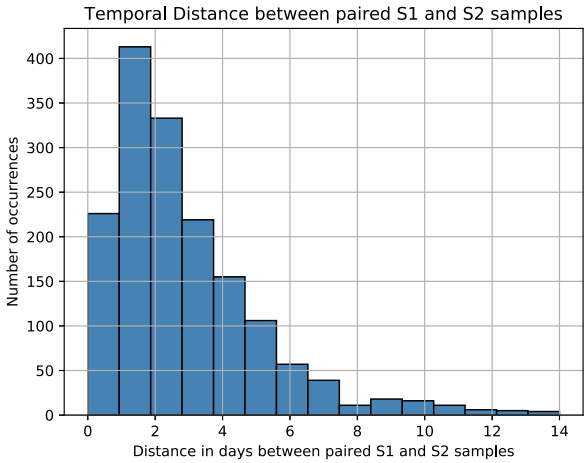


Fig. 12. Histogram of temporal differences between paired observations. The mean time differences across all paired observations are 2.61 ($\pm$ 2.41), indicating a close proximity between paired samples.

VI. Conclusion

As a large extent of our planet is covered by haze or clouds at any given point in time, such atmospheric distortions pose a severe constraint to the ongoing monitoring of earth. To approach this challenge, our work presented SEN12MS-CR-TS, a multimodal and multitemporal data set for training and evaluating global and all-season cloud removal methods. Our data set contains Sentinel-1 and Sentinel-2 observations from over 80000 km^2 of landcover, distributed globally and recorded through the year. The globally distributed ROIs are large-sized and capture a heterogeneous mass of landcover. We demonstrated the practicality of SEN12MS-CR by considering two methods: First, a model for sequence-to-point cloud removal. Second, a network for sequence-to-sequence cloud removal which, to our knowledge, provides the first case a model preserving temporal information is proposed in the context of cloud removal. Both methods benefited from the presence of coregistered and paired SAR measurements contained in our data set. The conducted experiments highlight the contribution of our curated data set to the remote-sensing community as well as the benefits of multimodal and multitemporal information to reconstruct noisy information. SEN12MS-CR is made public to facilitate future research in multimodal and multitemporal image reconstruction.

Appendix

Temporal Coincidence of Paired Observations

Full-scene observations of Sentinel-1 and Sentinel-2 are collected within a 14-day time window in a paired manner, as specified in Section II-A. To further analyze the temporal distance within paired data, Fig. 12 illustrates the empirically observed coincidences within SEN12MS-CR-TS. The mean time differences across all paired observations are 2.61 ($\pm$ 2.41), which is considerably smaller than the interval bound and implies a close proximity between paired samples.

ACKNOWLEDGMENT

The authors would like to thank ESA and the Copernicus program for making the Sentinel observations accessed for this submission publicly available. The authors would also like to thank Rewanth Ravindran for assisting us in the data curation process.

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] M. Schmitt and Y.-L. Wu, "Remote sensing image classification with the SEN12MS dataset," *ISPRS Ann. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. V-2-2021, pp. 101–106, Jun. 2021.

[3] M. U. Rafique, H. Blanton, and N. Jacobs, "Weakly supervised fusion of multiple overhead images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1479–1486.

[4] M. Schmitt, J. Prexl, P. Ebel, L. Liebel, and X. X. Zhu, "Weakly supervised semantic segmentation of satellite images for land cover mapping—Challenges and opportunities," 2020, *arXiv:2002.08254*.

[5] P. Ebel, S. Saha, and X. X. Zhu, "Fusing multi-modal data for supervised change detection," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. B3-2021, pp. 243–249, Jun. 2021.

[6] S. Saha, P. Ebel, and X. X. Zhu, "Self-supervised multisensor change detection," *IEEE Trans. Geosci. Remote Sens.*, early access, Sep. 15, 2022, doi: 10.1109/TGRS.2021.3109957.

[7] K. Enomoto *et al.*, "Filmy cloud removal on satellite imagery with multispectral conditional generative adversarial nets," 2017, *arXiv:1710.04835*.

[8] C. Grohnfeldi, 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. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2018, pp. 1726–1729.

[9] P. Singh and N. Komodakis, "Cloud-GAN: Cloud removal for Sentinel-2 imagery using a cyclic consistent generative adversarial networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2018, pp. 1772–1775. [Online]. Available: https://ieeexplore.ieee.org/document/8519033/

[10] 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.

[11] Z. Gu, Z. Zhan, Q. Yuan, and L. Yan, "Single remote sensing image dehazing using a prior-based dense attentive network," *Remote Sens.*, vol. 11, no. 24, p. 3008, Dec. 2019.

[12] V. Sarukkai, A. Jain, B. Uzkent, and S. Ermon, "Cloud removal in satellite images using spatiotemporal generative networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 1796–1805.

[13] 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.

[14] 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. 2020.

[15] 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.*, Sep. 2020, pp. 2065–2068.

[16] R. Bamler, "Principles of synthetic aperture radar," *Surv. Geophys.*, vol. 21, nos. 2–3, pp. 147–157, 2000.

[17] S. Oehmcke, T.-H.-K. Chen, A. V. Prishchepov, and F. Gieseke, "Creating cloud-free satellite imagery from image time series with deep learning," in *Proc. 9th ACM SIGSPATIAL Int. Workshop Anal. Big Geospatial Data*, Nov. 2020, pp. 1–10.

[18] Q. Zhang, Q. Yuan, Z. Li, F. Sun, and L. Zhang, "Combined deep prior with low-rank tensor SVD for thick cloud removal in multitemporal images," *ISPRS J. Photogramm. Remote Sens.*, vol. 177, pp. 161–173, Jul. 2021.

[19] A. Zupanc. (2017). *Improving Cloud Detection With Machine Learning*. Accessed: Oct. 10, 2019. [Online]. Available: https://medium.com/sentinel-hub/improving-cloud-detection-with-machine-learning-c09dc5d7cf13

[20] W. Sintarasirikulchai, T. Kasetkasem, T. Isshiki, T. Chanwimaluang, and P. Rakwatin, "A multi-temporal convolutional autoencoder neural network for cloud removal in remote sensing images," in *Proc. 15th Int. Conf. Electr. Engineering/Electronics, Comput., Telecommun. Inf. Technol. (ECTI-CON)*, Jul. 2018, pp. 360–363.

[21] K. Perlin, "Improving noise," in *Proc. 29th Annu. Conf. Comput. Graph. Interact. Techn.*, 2002, pp. 681–682.

[22] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore, "Google Earth engine: Planetary-scale geospatial analysis for everyone," *Remote Sens. Environ.*, vol. 202, pp. 18–27, Dec. 2017.

[23] L. Veci, P. Prats-Iraola, R. Scheiber, F. Collard, N. Fomferra, and M. Engdahl, "The Sentinel-1 toolbox," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Oct. 2014, pp. 1–3.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Oct. 2016, pp. 770–778.

[25] V. Lonjou *et al.*, "MACCS-ATCOR joint algorithm (MAJA)," *Proc. SPIE*, vol. 10001, Oct. 2016, Art. no. 1000107.

[26] J. H. Jeppesen, R. H. Jacobsen, F. Inceoglu, and T. S. Toftegaard, "A cloud detection algorithm for satellite imagery based on deep learning," *Remote Sens. Environ.*, vol. 229, pp. 247–259, Aug. 2019.

[27] D. López-Puigdollers, G. Mateo-García, and L. Gómez-Chova, "Benchmarking deep learning models for cloud detection in Landsat-8 and Sentinel-2 images," *Remote Sens.*, vol. 13, no. 5, p. 992, Mar. 2021.

[28] P. Ebel, M. Schmitt, and X. X. Zhu, "Internal learning for Sequence-to-Sequence cloud removal via synthetic aperture radar prior information," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2021, pp. 2691–2694.

[29] H. Zhang, L. Mai, N. Xu, Z. Wang, J. Collomosse, and H. Jin, "An internal learning approach to video inpainting," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 2720–2729.

[30] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham, Switzerland: Springer, 2015, pp. 234–241. [Online]. Available: https://link.springer.com/chapter/10.1007%2F978-3-319-24574-4_28

[31] M. Fuentes Reyes, S. Auer, N. Merkle, C. Henry, and M. Schmitt, "SAR-to-optical image translation based on conditional generative adversarial networks-optimization, opportunities and limits," *Remote Sens.*, vol. 11, no. 17, p. 2067, Sep. 2019.

[32] L. Wang, X. Xu, Y. Yu, R. Yang, R. Gui, Z. Xu, and F. Pu, "SAR-to-optical image translation using supervised cycle-consistent adversarial networks," *IEEE Access*, vol. 7, pp. 129136–129149, 2019.

[33] V. Lempitsky, A. Vedaldi, and D. Ulyanov, "Deep image prior," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9446–9454.

[34] F. De la Torre and M. J. Black, "Robust principal component analysis for computer vision," in *Proc. Int. Conf. Comput. Vis.*, vol. 1. Jul. 2001, pp. 362–369.

[35] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[36] X. X. Zhu and R. Bamler, "Tomographic SAR inversion by L_1-norm regularization-the compressive sensing approach," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 10, pp. 3839–3846, Feb. 2010.

[37] X. X. Zhu and R. Bamler, "A sparse image fusion algorithm with application to pan-sharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 5, pp. 2827–2836, May 2013.

[38] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[39] F. A. Kruse *et al.*, "The spectral image processing system (SIPS)-interactive visualization and analysis of imaging spectrometer data," *AIP Conf.*, vol. 283, no. 1, pp. 192–201, 1993.

[40] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[41] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Computer Vision—ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, pp. 694–711. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-46475-6_43

[42] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[43] M. Schmitt, L. Hughes, C. Qiu, and X. Zhu, "SEN12MS–A curated dataset of georeferenced multi-spectral Sentinel-1/2 imagery for deep learning and data fusion," *ISPRS Ann. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 4, pp. 153–160, Oct. 2019.

[44] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–15.
[45] M. Hintermüller and T. Wu, "Robust principal component pursuit via inexact alternating minimization on matrix manifolds," *J. Math. Imag. Vis.*, vol. 51, no. 3, pp. 361–377, 2015.
[46] S. S. Bucak, B. Gunsel, and O. Gursoy, "Incremental nonnegative matrix factorization for background modeling in surveillance video," in *Proc. IEEE 15th Signal Process. Commun. Appl.*, Oct. 2007, pp. 1–4.
[47] D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 13, 2000, pp. 556–562. [Online]. Available: https://papers.nips.cc/paper/2000/hash/f9d1152547c0bde01830b7e8bd60024c-Abstract.html
[48] N. Guan, D. Tao, Z. Luo, and J. Shawe-Taylor, "MahNMF: Manhattan non-negative matrix factorization," 2012, *arXiv:1207.3438*.
[49] A. Sobral, S. Javed, S. K. Jung, T. Bouwmans, and E.-H. Zahzah, "Online stochastic tensor decomposition for background subtraction in multispectral video sequences," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 106–113.

Patrick Ebel received the B.Sc. degree in cognitive science from the University of Osnabrück, Osnabrück, Germany, in 2015, and the M.Sc. degree in cognitive neuroscience and the M.Sc. degree in artificial intelligence from Radboud University Nijmegen, Nijmegen, The Netherlands, in 2018. He is currently pursuing the Ph.D. degree with the Data Science in Earth Observation Laboratory, Department of Aerospace and Geodesy, Technical University of Munich (TUM), Munich, Germany.

His research interests include machine learning as well as its applications in computer vision and remote sensing. Specifically, he is working on multimodal and multitemporal data fusion and automated image reconstruction methods.

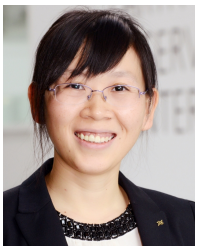
Yajin Xu received the B.Sc. degree (Hons.) in engineering from Wuhan University, Wuhan, China, in 2018. He is currently pursuing the joint M.Sc. degree in earth-oriented space science and technology (ESPACE) with Wuhan University and the Technical University of Munich (TUM), Munich, Germany.

His study focus is machine learning applied to remote-sensing data. In 2021, he was a Research Assistant with the Data Science in Earth Observation (SiPEO) Group, TUM and Remote Sensing Technology Institute of DLR, investigating deep learning-based approaches for cloud removal in optical satellite data. His interest includes geospatial data analysis.

Michael Schmitt (Senior Member, IEEE) received the Dipl.-Ing. (Univ.) degree in geodesy and geoinformation, the Dr.-Ing. degree in remote sensing, and the Habilitation degree in data fusion from the Technical University of Munich (TUM), Munich, Germany, in 2009, 2014, and 2018, respectively.

Since 2021, he has been the Chair of Earth Observation at the Department of Aerospace Engineering, Bundeswehr University Munich, Neubiberg, Germany. Before that, he was a Full Professor of applied geodesy and remote sensing with the Department of Geoinformatics, Munich University of Applied Sciences, Munich. From 2015 to 2020, he was a Senior Researcher and the Deputy Head at the Professorship for Data Science in Earth Observation at TUM. In 2019, he was additionally appointed as an Adjunct Teaching Professor with the Department of Aerospace and Geodesy, TUM. In 2016, he was a Guest Scientist at the University of Massachusetts, Amherst, MA, USA. His research focuses on image analysis and machine learning applied to the extraction of information from multimodal remote sensing observations. In particular, he is interested in remote-sensing data fusion with a focus on SAR and optical data.

Dr. Schmitt is a Co-Chair of the Working Group "SAR and Microwave Sensing" of the International Society for Photogrammetry and Remote Sensing and also the Working Group "Benchmarking" of the IEEE-Geoscience and Remote Sensing Society (GRSS) Image Analysis and Data Fusion Technical Committee. He frequently serves as a reviewer for a number of renowned international journals and conferences and has received several best reviewer awards. He is an Associate Editor of IEEE GEOSCIENCE AND REMOTE SENSING LETTERS.



Xiao Xiang Zhu (Fellow, IEEE) received the M.Sc., Dr.-Ing., and Habilitation degrees in signal processing from the Technical University of Munich (TUM), Munich, Germany, in 2008, 2011, and 2013, respectively.

She is a Professor of data science in earth observation (former: signal processing in earth observation) at TUM and the Head of the Department "EO Data Science" at the Remote Sensing Technology Institute, German Aerospace Center (DLR). Since 2019, she has been a Co-Coordinator of the Munich Data Science Research School and also the Head of the Helmholtz Artificial Intelligence—Research Field "Aeronautics, Space and Transport." Since May 2020, she has been the Director of the International Future AI Laboratory "AI4EO—Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond," Munich. Since October 2020, she has also been the Co-Director of the Munich Data Science Institute (MDSI), TUM. She was a Guest Scientist or a Visiting Professor at the Italian National Research Council (CNR-IREA), Naples, Italy; Fudan University, Shanghai, China; the University of Tokyo, Tokyo, Japan; and the University of California, Los Angeles, CA, USA; in 2009, 2014, 2015, and 2016, respectively. She is currently a Visiting AI Professor at ESA's Phi-Laboratory, Frascati, Italy. Her main research interests are remote sensing and earth observation, signal processing, machine learning, and data science, with a special application focus on global urban mapping.

Dr. Zhu is a member of young academy (Junge Akademie/Junges Kolleg) at the Berlin-Brandenburg Academy of Sciences and Humanities and the German National Academy of Sciences Leopoldina and the Bavarian Academy of Sciences and Humanities. She serves in the Scientific Advisory Board in several research organizations, among others the German Research Center for Geosciences (GFZ) and the Potsdam Institute for Climate Impact Research (PIK). She is an Associate Editor of IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING and serves as the Area Editor responsible for special issues of *IEEE Signal Processing Magazine*.