

COMPARISON OF CONVOLUTIONAL NEURAL NETWORKS FOR CLOUDY OPTICAL IMAGES RECONSTRUCTION FROM SINGLE OR MULTITEMPORAL JOINT SAR AND OPTICAL IMAGES

Rémi Cresson^{a,*}, Nicolas Narçon^a, Raffaele Gaetano^b, Aurore Dupuis^c, Yannick Tanguy^c, Stéphane May^c, and Benjamin Commandré^a

(a) INRAE, 500 rue Jean-François Breton, 34090 Montpellier, France

(b) CIRAD, 500 rue Jean-François Breton, 34090 Montpellier, France

(c) CNES, 18 Av. Edouard Belin, 31400 Toulouse, France

KEY WORDS: Image reconstruction, clouds, atmospheric perturbation, SAR, optical, convolutional neural networks, gap-filling

ABSTRACT:

With the increasing availability of optical and synthetic aperture radar (SAR) images thanks to the Sentinel constellation, and the explosion of deep learning, new methods have emerged in recent years to tackle the reconstruction of optical images that are impacted by clouds. In this paper, we focus on the evaluation of convolutional neural networks that use jointly SAR and optical images to retrieve the missing contents in one single polluted optical image. We propose a simple framework that ease the creation of datasets for the training of deep nets targeting optical image reconstruction, and for the validation of machine learning based or deterministic approaches. These methods are quite different in terms of input images constraints, and comparing them is a problematic task not addressed in the literature. We show how space partitioning data structures help to query samples in terms of cloud coverage, relative acquisition date, pixel validity and relative proximity between SAR and optical images. We generate several datasets to compare the reconstructed images from networks that use a single pair of SAR and optical image, versus networks that use multiple pairs, and a traditional deterministic approach performing interpolation in temporal domain.

1. INTRODUCTION

1.1 Context

The Sentinel constellation is composed of different coupled SAR and optical sensors with short revisit period (five to ten days). However, optical images are frequently polluted by cloud cover. To leverage the problem of optical image reconstruction, various approaches have been proposed over the years. First, approaches based on mathematical, physical or statistical model, have been extensively used to reconstruct the missing parts of the images. A review of these traditional approaches are summarized in (Shen et al., 2015). Among them, we can distinguish multispectral based (Hu et al., 2015) methods, multitemporal (Cheng et al., 2014) (Li et al., 2014), and methods using optical and SAR data fusion (Eckardt et al., 2013). Lastly, machine learning and particularly deep learning have become popular to achieve the task of cloudy images reconstruction, thanks to the unprecedented ability to fuse images of different modalities, and accompanied with state of the art results. In recent years, deep neural networks have proven to be effective for image reconstruction from time series of same modality (Zhang et al., 2018) or from timely available images at coarser spatial resolution (Liu et al., 2019), or from joint optical and SAR time series (Scarpa et al., 2018) (Cresson et al., 2019). (Sarukkai et al., 2020) have casted the problem of cloud removal as a conditional image synthesis challenge, and have proposed a network to remove clouds from a single optical image or from a triplet of optical images. In (Meraner et al., 2020), a cloudy optical image is reconstructed with the help of a single SAR, using a convolutional neural network with a residual correction performed on the input cloudy optical image. Conversely, (Ebel et al., 2020) have jointly modeled the

cloud removal and the synthetic cloudy optical images generation problems, concluding that the networks trained over real data were performing the best.

1.2 Problematic

Our purpose is to lead the evaluation of several approaches based on convolutional neural networks trained on real data, that reconstruct optical images impacted by clouds. We compare these approaches with a popular traditional deterministic approach, the gap-filling (Inglada et al., 2015). While the existing literature covers mostly the comparison of approaches that consume the same inputs, comparing approaches consuming various forms of inputs, e.g. single or multiple, optical and/or SAR images or pairs of images, remains an interesting topic from an operational perspective. In this paper, we address the comparison of the following kinds of methods, which consume one or more input pair of optical and SAR images acquired in various conditions to reconstruct or generate one single output optical image:

1. Reconstruct a cloudy optical image using an additional SAR image acquired at the same date,
2. Reconstruct a cloudy optical image using an additional SAR image acquired at the same date, and two other cloudy optical/SAR images pairs acquired before and after,
3. Generate an optical image at one given desired date using two clean optical images acquired before and after.

1.3 Method

In this paper, we provide an insight into various optical image reconstruction methods. In particular, we address the question

* Corresponding author

of which approach to employ for a specific availability of remote sensing products, in comparing a few selected single date based and multiple dates based methods. Since these methods employ inputs of different nature (single optical or SAR image, or optical and SAR images pair), number (single or multiple image or pair or images), and cloud coverage (clean or cloudy images), their comparison is not straightforward. To leverage this, we introduce the *acquisitions layout*, a descriptor of the available inputs and their properties, for a specific approach. We then use space partitioning data structures to ease the generation of various datasets from specific acquisitions layouts. These datasets are then used to train networks, and also at inference time for the comparison of the different methods when a common set of inputs can be shared and matches the expected constraints, e.g. cloud coverage. This simple yet generic framework allows to produce datasets tailored for a specific problem and suited to the data availability, i.e. inputs and targets images. We carry out the benchmarks of representative state of the art methods for optical image reconstruction, namely the network presented in (Meraner et al., 2020), which uses a single pair of optical and SAR image acquired near the same date, and a convolutional network that inputs three consecutive pairs to reconstruct the central optical image (Cresson et al., 2019)(Scarpa et al., 2018). In (Scarpa et al., 2018), an additional DEM is used as input of the network, and we also investigate the contribution of such ancillary data in the single date network. To better disentangle the benefits of the different modalities, we perform an ablation study removing the DEM, and the SAR inputs.

1.4 Overview

In section 2, we present the remote sensing data used in this study. In section 3, we detail the implemented models. In section 4, we detail our framework for the creation of datasets, which is a crucial aspect of our work. In section 5 we detail the methodology used to train the models and carry out the comparison of the different approaches. Finally we discuss the results in section 6.

2. DATA

2.1 Sentinel images

We use 10 tiles of Sentinel-2 images acquired over the Occitanie area in France (figure 1), from January 2017 to January 2021, that represents a total of 3593 optical images. We also use every available Sentinel-1 images acquired in ascending orbit over the Occitanie area during the same period, that we superimpose over the Sentinel-2 images pixels grids (more details are provided in section 2.1.1), which represents a total of 5136 SAR images. We believe that since a large part of the earth is covered only with single orbit (i.e. ascending or descending), our study results would be more easily reproducible with a single orbit for SAR images, hence we use only the ascending orbit over our study area. Table 1 summarizes the number of Sentinel images used for this study. The total area covered by the Sentinel tiles is $106.7 \times 10^3 km^2$. The following sections details the Sentinel-1 and Sentinel-2 products.

2.1.1 Sentinel-1 images We have used the so-called S1Tiling tool¹ to automatically download and process the Sentinel-1 images. The tool performs the orthorectification and

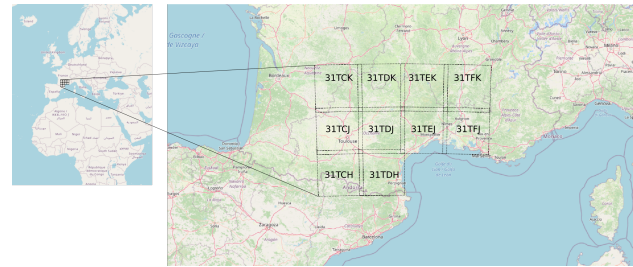


Figure 1. The region of interest, located in the Occitanie area (south of France mainland). Sentinel-2 images envelopes are plotted in black. Map data © OpenStreetMap contributors, CC BY-SA

the calibration in sigma nought of the VV and VH SAR images channels. It also projects and resamples the final images over the same coordinate reference system and pixel grid as the Sentinel-2 images, at 10m physical spacing resolution.

2.1.2 Sentinel-2 images The Theia Land data center² provides Sentinel-2 images in surface reflectance. The products are computed using MACCS (Multi-sensor Atmospheric Correction and Cloud Screening), a level 2A processor which detects the clouds and their shadows, and estimates aerosol optical thickness, water vapour and corrects for the atmospheric effects (Hagolle et al., 2015). While level 1C processing level could have been used in this study, we chose level 2A products because they include cloud masks that are useful meta-data. Indeed these information suffice to derive a cloud coverage percentage over patches. Figure 2 shows the cloud coverage computed for each location from the number of cloudy pixels among available ones in the temporal dimension. The average cloud coverage over the area is 39.1% and the standard deviation 6.8%. An evaluation of the cloud masks is provided in (Baetens et al., 2019). To discharge storage and computational requirements, we used only the 10m spacing bands, i.e. spectral bands number 2, 3, 4 and 8.

Tile	S1	S2
31TCH	684	374
31TCJ	456	380
31TCK	454	273
31TDH	462	370
31TDJ	462	302
31TDK	462	286
31TEJ	672	512
31TEK	595	311
31TFJ	443	408
31TFK	446	377

Table 1. Number of Sentinel-1 (S1) and Sentinel-2 (S2) images used over the Occitanie area (France)

2.1.3 Digital Elevation Model The Digital Elevation Model (DEM) from the Shuttle Radar Topography Mission (SRTM)(Farr and Kobrick, 2000) is used to perform the processing of Sentinel-1 images. It is also used in our $SSOP_{unet+DEM}$ and $MSOP_{unet+DEM}$ networks, introduced in the following sections.

3. ARCHITECTURES

The implemented deep learning based models are detailed in the subsections below. The number of trainable parameters, batch

¹ <https://gitlab.orfeo-toolbox.org/s1-tiling/s1tiling>

² <https://www.theia-land.fr/en/product/sentinel-2-surface-reflectance/>

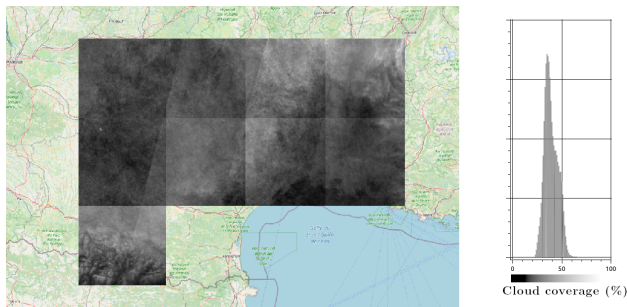


Figure 2. The cloud coverage over the area. Left: the percentage of cloud-free pixels in computed at each location. Right: histogram of the values. Mean cloud coverage: 39.13%, standard deviation: 6.80%. Map data © OpenStreetMap contributors, CC BY-SA

size used per GPU, and training time is summarized in table 2 for each model.

Model	Nb. params	Batch	Training time
$SSOP_{mer}$	18,905M	8	3136h.GPU
$SSOP_{unet}$	16.652M	64	104h.GPU
$MSOP_{unet}$	29.272M	32	149h.GPU

Table 2. Number of parameters in the trained models, batch size used, and training time (in hours.GPUs) for the training step on a NVIDIA V100 GPU with 32Gb RAM. The training time corresponds to the duration required to computed the best model.

3.1 Single SAR/Optical pair (SSOP)

We denote $SSOP$, the approaches that input one SAR image and one optical image polluted by cloud, and which reconstruct the missing parts of the optical image. Introduced in (Meraner et al., 2020), this kind of approach is trained and evaluated from samples composed of $(S1_t, S2_t, S2_{t'})$ triplets, where $S2_t$ is an optical image potentially polluted by clouds, with $S1_t$ and $S2_t$ acquired close together and $S2_{t'}$ a cloud-free optical image acquired close to the $(S1_t, S2_t)$ pair. Figure 3 illustrates the architecture of this family of networks.

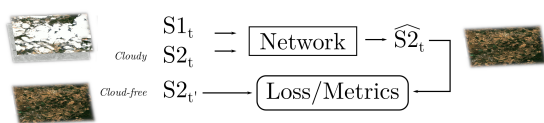


Figure 3. SSOP network. $S1_t$ and $S2_t$ denotes the input pair of Optical and SAR images acquired at date t . The $S2_{t'}$ denotes the reference image used to compute the loss (during training) and the metrics (at inference time from test data) from the reconstructed optical image $\hat{S2}_t$.

3.1.1 $SSOP_{mer}$ We implement the network described in (Meraner et al., 2020), which uses a residual correction from a ResNet backbone (He et al., 2016) to reconstruct the output optical image from a pair of one SAR image and one cloudy optical image. However, the authors of (Meraner et al., 2020) did not have $S2_{t'}$ acquisitions systematically close to $S2_t$, therefore they use an additional loss based on cloud masks to encourage the identity transformation of cloud-free pixels. Since

our goal is to use a simple information about the presence of clouds, namely an approximation of the cloud cover percentage in a Sentinel-2 image or a set of patches, we don't use such pixel-wise cloud-mask based loss. Moreover, as we control the gathering of samples matching the acquisitions layout described in table 4, a cloud-free $S2_{t'}$ image acquired close to $S2_t$ is always available. We hypothesize that the contribution of such loss is likely marginal since none or very little changes should happen between $S2_{t'}$ and $S2_t$. Hence we train the network using only the l_1 loss. We denote $SSOP_{mer}$ the implementation of this network.

3.1.2 $SSOP_{unet}$ We implement a modified version of the previously described architecture, employing a U-Net backbone (Ronneberger et al., 2015) instead of ResNet. Our motivation behind this modification is that the ResNet backbone has two disadvantages compared to U-Net: (i) convolutions are applied without any stride, which consumes a lot more memory and requires much more operations since all convolutions apply over the entire images at their native resolution i.e. without any downsampling of the features, (ii) all inputs have to be resampled at the smallest input images resolution, i.e. a physical spacing of 10 m, since all the network features are computed at the same resolution and no downscaling is performed across the network, which is computationally counterproductive. We denote $SSOP_{unet}$ our implementation of this modified network. To illustrate the advantage of additional inputs at a lower resolution, we use an additional Digital Elevation Model (DEM) as input, resampled at 20m, as shown in figure 4. Our modified network is illustrated in figure 5. We denote $SSOP_{unet+DEM}$ the model with the input DEM. It can be noted that such a model could also generate outputs of different resolutions, typically the Sentinel-2 spectral bands at 20m.

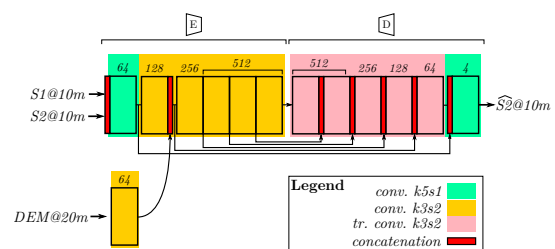


Figure 4. Proposed architecture for encoder (E) and decoder (D), enabling the use of inputs of different resolutions. First and last convolutions use a unitary stride and a kernel of size 5. Other convolutions use strides 2 and a kernel of size 3. Skip connections between the encoder and the decoder perform the concatenation of the features from the encoder with the decoder outputs. All convolutions except the last $k5s1$ convolution are followed with a ReLU activation function. No batch normalization is used.

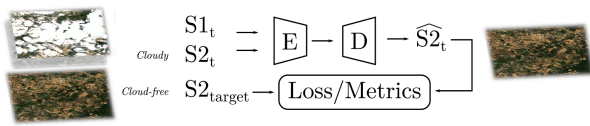


Figure 5. Our modified SSOP network. $S1_t$ and $S2_t$ denotes the input pair of Optical and SAR images. The $S2_{t'}$ denotes the reference image used to compute the loss (during training) and the metrics (at inference time from test data) from the reconstructed optical image $\widehat{S2}_t$. (E) and (D) denotes respectively the encoder and the decoder of the U-Net backbone.

3.2 Multiple SAR/Optical pairs (MSOP)

A number of approaches using multiple pairs of optical and SAR images have been presented in the literature. For instance, (Scarpa et al., 2018) use two pairs of SAR/optical acquired before and after date t , a Digital Elevation model (DEM), and an additional SAR image acquired at date t to estimate radiometric indices at date t . Conversely, (Cresson et al., 2019) use multiple optical and SAR images to generate a synthetic optical image at date t . While these works were carried on input cloud-free images, a similar network architecture can also be applied on cloudy input images to retrieve the missing contents of the optical image at date t .

3.2.1 $MSOP_{UNET}$ We build a multitemporal network inspired from the architectures presented in (Scarpa et al., 2018) and (Cresson et al., 2019). We generalize to multitemporal the approach of (Meraner et al., 2020) with a new architecture that inputs multiple SAR/optical images pairs at $t-1$, t and $t+1$ and a DEM, aiming to reconstruct the potentially damaged optical image at t . We use a similar architecture as the encoder/decoder U-Net backbone of the $SSOP_{UNET}$ model, except that encoder weights are shared for $t-1$, t and $t+1$ inputs, and features from three encoders (E) are concatenated before being processed with the decoder (D), which outputs the reconstructed optical image at t . Unlike (Meraner et al., 2020), our model does not employ residual connections to generate the reconstructed optical image (figure 6).

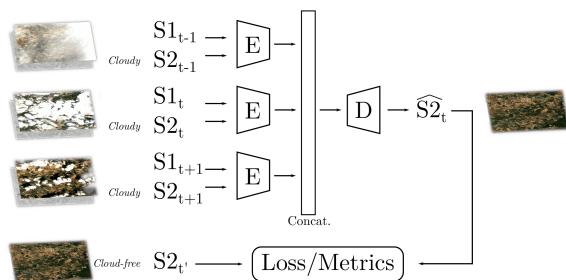


Figure 6. $MSOP_{UNET}$ network. $(S1_{t-1}, S2_{t-1})$, $(S1_t, S2_t)$ and $(S1_{t+1}, S2_{t+1})$ denotes the input pairs of optical and SAR images. $S2_{t'}$ denotes the reference image used to compute the loss (during training) and the metrics (at inference time from test data) from the reconstructed optical image $\widehat{S2}_t$.

We denote $MSOP_{UNET+DEM}$ the MSOP model using the input DEM in (E) as shown in figure 4.

3.3 Gap-filling

The *Gap-filling* consists in interpolating temporally close optical images to approximate one target image (Inglada et al., 2015). While gap-filling is not a reconstruction method, i.e. the input image at t is not used, it is commonly used as such, in estimating the parts of the image that is polluted by clouds. Gap-filling is restricted to cloud-free input images, and do not use SAR images. In the case of a linear model, the generated output image $\widehat{S2}_t$ can be written using the following formula:

$$\widehat{S2}_t = S2_{t-1} + (S2_{t+1} - S2_{t-1}) \times \frac{T_t - T_{t-1}}{T_{t+1} - T_{t-1}} \quad (1)$$

Where T is the timestamp, in seconds, of the dates.

4. DATASETS

4.1 Acquisitions layouts

The so-called *Acquisitions layout* describes inputs and targets of a specific use-case scenario. In the particular case of image reconstruction addressed in this paper, the acquisitions layouts presented in the following sections have one common item, namely $S2_{t'}$ the target cloud-free optical image. Depending on the approaches, the acquisitions layout can include additional items, for instance:

- A single SAR image
- A single optical image
- A pair of SAR + optical image

For each item, the acquisitions layout describes crucial properties:

- For each optical image: a range of cloud coverage percentage (e.g. $[0, 10]$),
- For each (SAR, optical) pair: the maximum temporal gap between the two images acquisition dates, in hours,
- For each SAR or optical image: the acquisition date range, relative to a reference item of the acquisitions layout (e.g. $[240h, 360h]$).

This simple yet generic description formalizes of how the images are acquired for a particular use-case scenario. We have carefully crafted acquisitions layouts that represents the operational context of use of the approaches, i.e. for which it is possible to use them on every available images. For instance, to chose the maximum temporal gap between the SAR and the optical images acquisition dates, we have analyzed the distribution of the temporal gap between the closest (S1, S2) images (figure 7). Since more than 96% of the nearest (S1, S2) pairs are close to 72 hours, we used this duration as the maximum temporal gap in SAR-optical images pairs. We provide in section 4 all acquisitions layouts suited for the training and testing of the involved networks, and explain how the other parameters (i.e. time ranges for each acquisition layout items) are chosen.

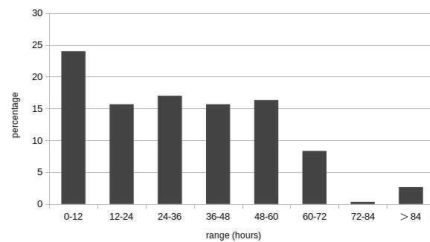


Figure 7. Distribution of the temporal gap between the closest (S1, S2) images from the available images.

4.2 Patches indexation

Figure 8 gives an overview of the dataset creation step. In order to perform the query of samples, i.e. the search of groups of patches that match the properties defined in the acquisitions layout, we use an indexation structure. An R-Tree indexes all available patches. The space partitioning data structure describes the following dimensions:

- Cloud coverage (ranging from 0% to 100%)
- Time from the reference image of the acquisitions layout
- Duration to the closest valid SAR patch
- Number of pixels different from the no-data value

A Kd-Tree is used to ease the computation of the duration to the closest valid SAR patch. The R-Tree is built for each patches of Sentinel tiles using the acquisition dates provided in the Sentinel images metadata, and the following statistics collected on the Sentinel images patches:

- **For Sentinel-1 images:** the number of valid pixels,
- **For Sentinel-2 images:** the number of valid pixels, and the number of pixels impacted by clouds. To compute this last, we use the cloud quality mask provided in the Theia product.

We note that cloud masks are only used as a single value for each patch, representing the proportion of cloud coverage. We have computed the R-Trees from non-overlapping, 256×256 sized patches in all Sentinel tiles. This indexation structure is computed once. After that, any acquisitions layout can be used to query all samples matching the defined properties of the remote sensing acquisitions. One generated sample includes data arrays containing pixels and ancillary data, e.g. acquisition date, for each items of the acquisitions layout. Finally, the samples are restricted in the provided region of interest, to allow the generation of mutually exclusive samples in the geographical domain, i.e. in training, validation and test datasets.

4.3 Generation of samples

For each network, samples are first extracted in the images from the specific acquisitions layout. We split the samples in three groups: training, validation, and test. We ensure that these three groups are mutually exclusive by randomly selecting their location in the geographical domain, without overlapping patches of distinct groups. We have randomly selected 5% and 15% of the area to form the region of interest for the validation and test

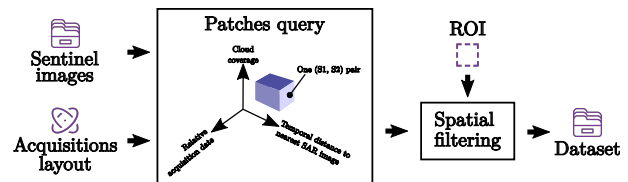


Figure 8. Workflow for the creation of datasets. The sample query use a R-Tree indexing the available Sentinel-1 and Sentinel-2 images patches.

datasets, and the other 80% has been used to form training datasets. Since the swath of Sentinel-1 and 2 does overlap in some areas, the samples density is heterogeneous in spatial domain. For this reason, we have limited the number of samples per spatial location in the training and validation datasets, to guarantee that models are trained with the same number of patches at each location. Thus for the training and validation datasets, a maximum amount of 50 samples has been collected at each locations. For the test dataset, all available samples are extracted. Table 3 summarize the number of distinct samples for training, validation and test datasets. The differences in samples number is due to the availability of images, or pairs of images, depending on the properties defined in the acquisitions layout, i.e. number and type of acquisitions, cloud coverage and temporal constraints. All acquisitions layouts used to generate the datasets are detailed in the following subsections.

Dataset	Training	Validation	Test
SSOP	600.1k	35.3k	70.6k
MSOP	600.1k	35.3k	70.6k
MSOP _{cld}	/	/	76k

Table 3. Number of samples in each datasets. The SSOP, MSOP and MSOP_{cld} datasets are used respectively to train the mono-temporal networks, the multi-temporal networks, and to compare all models with the gap-filling over an acquisitions layout matching all approaches validity domains.

In the following sections, we detail the properties of each acquisitions layout of the datasets.

4.3.1 SSOP (Single date inputs cloudy optical image)

The acquisitions layout for SSOP networks training is presented in table 4. We chose the following parameters for the acquisitions layout: the maximum gap between the acquisition dates of S1 and S2 images is set to 72 hours and the maximum spread between the cloud-free and the polluted optical images is set to 10 days. With this settings, we reach a total number of 600k samples for training, which is approximately 4 times the amount of samples used in the original paper of (Meraner et al., 2020).

Name	S1 ($\pm \Delta t$)	S2 (% clouds)	Time-stamp
t	Yes ($\pm 72h$)	[0, 100]	Reference
t'	/	0	$[-10d, +10d]$

Table 4. Acquisitions layout for the SSOP networks, depicting the $(S1_t, S2_t)$ images pair acquired close to a single cloud-free optical image $(S2_{t'})$.

4.3.2 MSOP (Multitemporal inputs, any optical images i.e. cloudy or not)

The acquisitions layout for the training of the

Name	S1 ($\pm\Delta t$)	S2 (% clouds)	Time-stamp
$t - 1$	Yes ($\pm 72h$)	0	$[-18d, -10d]$
t	Yes ($\pm 72h$)	100	reference
t'	/	0	$[-5d, +5d]$
$t + 1$	Yes ($\pm 72h$)	0	$[+10d, +18d]$

Table 6. Acquisitions layout enabling the comparison of the gapfilling and the $MSOP_{unet}$ network. The cloud coverage at t is 100%, all other optical images remain cloud free.

$MSOP_{unet}$ network is presented in table 5. It consists of three optical images at $t - 1$, t and $t + 1$ that can be polluted by clouds, and one cloud-free optical image at t' , used as the training target. As explained in section 4.1, we used 72 hours for the maximum gap between the acquisition dates of S1 and S2 images. The cloud-free optical image is acquired at most 10 days from the optical image at t , to roughly falls within the less frequent revisit cycle of the Sentinel-2 constellation everywhere over our study area. Finally, we have selected a temporal range for $t - 1$ and $t + 1$ dates that avoids the cloud-free optical image acquisition date, and that also falls within the revisit cycle of the Sentinel-2 constellation, i.e. 10 to 18 days.

Name	S1 ($\pm\Delta t$)	S2 (% clouds)	Time-stamp
$t - 1$	Yes ($\pm 72h$)	$[0, 100]$	$[-18d, -10d]$
t	Yes ($\pm 72h$)	$[0, 100]$	reference
t'	/	0	$[-10d, +10d]$
$t + 1$	Yes ($\pm 72h$)	$[0, 100]$	$[+10d, +18d]$

Table 5. Acquisitions layout used to train the $MSOP_{unet}$ network.

4.4 MSOP_{old} (Multitemporal inputs, with cloud-free optical images at $t - 1$ and $t + 1$, and one cloudy image at t)

Table 6 shows one acquisitions layout enabling the comparison of the gap-filling with the SSOP and MSOP networks, thanks to cloud-free optical images available at $t - 1$ and $t + 1$, and one completely cloudy optical images at t , which intends to make as fair as possible the comparison. We denote the corresponding dataset MSOP_{old}. In this acquisitions layout, the cloud-free optical image at t' , acquired at most 5 days from the date t , is used to compute the metrics over the reconstructed image.

We note that our settings make possible the extraction of a sufficient number of samples, thanks to the availability of Sentinel-1 and Sentinel-2 over our study site, but this setting might be adjusted for other regions of the world where the Sentinel coverage is less timely available.

5. BENCHMARKS

5.1 Protocol

We train all networks with their respective datasets presented in section 2. The SSOP and MSOP models are trained over the training dataset detailed in section 4.3.1, and evaluated over the test datasets detailed in sections 4.3.2 and 4.4. The MSOP models are trained over the dataset detailed in section 4.3.2, and evaluated on the test datasets detailed in sections 4.3.2 and 4.4. We use the ADAM algorithm (Kingma and Ba, 2014) to train all networks to minimize the l_1 loss. For $MSOP_{unet}$ and

$SSOP_{unet}$, we use a learning rate of $lr = 0.00012$, $\beta_0 = 0.9$ and $\beta_1 = 0.999$, with a batch of size 128 distributed across 4 GPUs. We train the $SSOP_{mer}$ network using the same setup as described by the authors. All experiments are realized on NVIDIA V100 GPUs with 32Gb RAM. We kept the trained models that reach the smallest l_1 loss on the validation split. To assess the performance of each approach, we compute the following metrics between the reconstructed output optical images $\widehat{S2}_t$ and the reference cloud-free optical image $S2_{t'}$ over the test datasets:

- Peak Signal to Noise Ratio (PSNR):

$$PSNR = 10 \times \log_{10} \left(\frac{d^2}{MSE} \right) \quad (2)$$

Where MSE is the Mean Squared Error computed over n patches:

$$MSE = \frac{1}{n} \sum \|\widehat{S2}_t - S2_{t'}\|^2 \quad (3)$$

The higher is the PSNR, the closer are the values of the estimated image to the target image.

- Spectral angle (SAM)(Kruse et al., 1993), representing the mean spectral angle between the estimated image and the target image, ranging in $[0, \pi]$
- The Structural Similarity Index (SSIM)(Wang et al., 2004), measures the similarity in terms of structure, between the estimated image and the target image. The range is $[0, 1]$, and values close to 1 correspond to the best structural similarity between the images.

5.2 Ablation study

In order to assess the SAR, DEM, and optical modalities benefits, we have performed an ablation study. For the SSOP and MSOP networks, we have derived two modified architectures, one without the SAR input, and the other without DEM and without SAR. For a sake of computational budget, we only have studied the ablation with the U-Net based networks.

5.3 Results

In the following section, we report the evaluation metrics computed on all test datasets.

5.3.1 Comparison of SSOP networks The comparison between SSOP networks is carried out on the test dataset presented in section 4.3.1. We first report the metrics obtained with SSOP networks in table 7. It can be noticed that the metrics from the modified network ($SSOP_{unet}$) are close to the original $SSOP_{mer}$. SSIM and SAM are slightly better for $SSOP_{mer}$ and MSE and PSNR a bit better for $SSOP_{unet}$. While this result is not groundbreaking in terms of evaluation metrics, we highlight the huge difference of required overall processing time: to train both networks over the same dataset with the same setup, $SSOP_{unet}$ needs ≈ 30 hours and $SSOP_{mer} \approx 35$ days. For this particular reason, we have chosen to perform all other benchmarks only on $SSOP_{unet}$, since the metrics are quite similar to $SSOP_{mer}$, but the processing budget far lower and we could ran more experiment at lower cost. It can be noticed that the $SSOP_{unet+DEM}$ model has the best PSNR, MSE and SAM, but the $SSOP_{mer}$ still has a slightly better SSIM. One explanation could be that in the

$SSOP_{mer}$ model, all convolutions are performed in the original resolution, and no downsampling is performed, which might preserve the structural similarity, hence a better SSIM. The lowest metrics are obtained with the $SSOP_{UNET,w/oSAR}$ model, which does not use DEM and SAR inputs, showing the benefits of these modalities for the reconstruction.

Model	MSE	SSIM	PSNR	SAM
$SSOP_{UNET,w/oSAR}$	324508	0.8388	24.888	0.1595
$SSOP_{mer}$	277971	0.8656	25.560	0.1425
$SSOP_{UNET}$	261223	0.8568	25.830	0.1448
$SSOP_{UNET+DEM}$	234410	0.8645	26.300	0.1401

Table 7. Comparison of SSOP networks performed on the SSOP dataset

Figure 9 shows images from the test dataset, processed with the different U-Net based SSOP networks. We can visually appreciate the contributions of the input SAR and DEM. It can be noticed the limits of the method with thick clouds in the optical image, especially for the $SSOP_{UNET,w/oSAR}$ network that only use the input optical image. Figure 10 show the limits of the $SSOP_{UNET+DEM}$ network with very thick atmospheric perturbation.

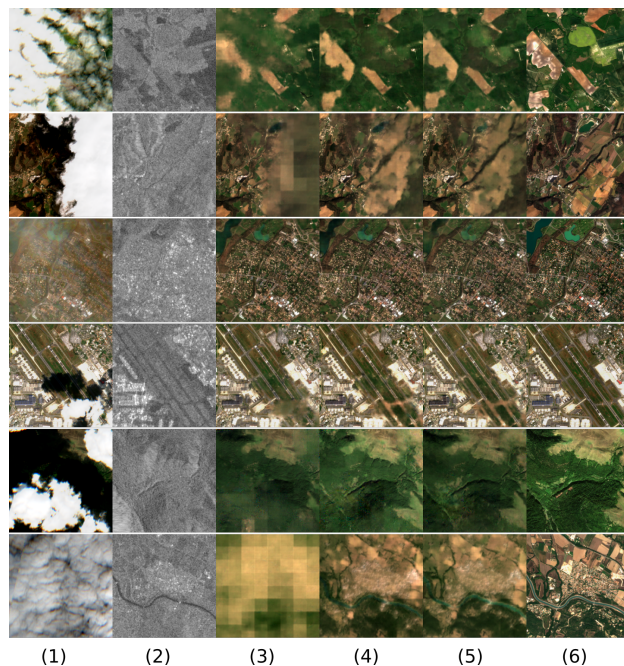


Figure 9. From left to right: input cloudy optical image $S2_t$ (1), input SAR image $S1_t$ (2), output $\hat{S2}_t$ from $SSOP_{UNET,w/oSAR}$ (3), $SSOP_{UNET}$ (4), and $SSOP_{UNET+DEM}$ (5), (6) reference image

5.4 Comparison of SSOP and MSOP networks

The comparison of approaches that input one or more cloudy images to reconstruct the optical image at t , is carried out. We compare networks that consume different kind of inputs, i.e. one single ($S1$, $S2$) pair for SSOP networks versus three pairs of images for MSOP networks. We recall that, unlike the MSOP networks, $t-1$ and $t+1$ images are not used by the SSOP networks. We compare the networks on the test dataset detailed in section 4.3.2, since its acquisitions layout fulfills both MSOP

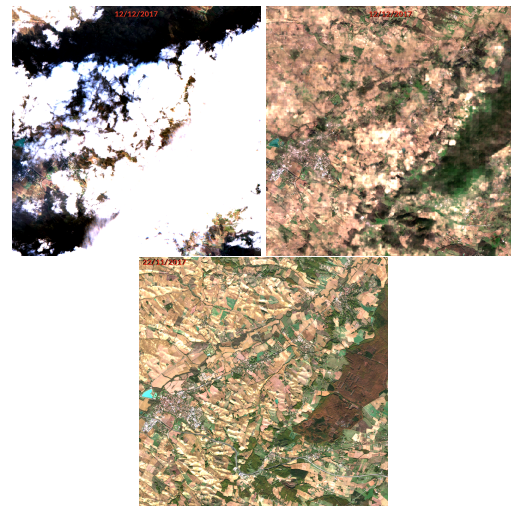


Figure 10. Limits of the SSOP networks with thick clouds in optical images. Top left: input cloudy optical image $S2_t$, Top right: SSOP network output $\hat{S2}_t$. Bottom: the reference image $S2_{t'}$.

and SSOP models validity domains, in particular the maximum SAR-optical temporal gap at t . Evaluation metrics are reported in table 8. While it can be observed the same outcome in the comparison between SSOP networks, these quality metrics differ a bit from the ones presented in table 7, since the evaluated samples are just a subset of this last dataset. The qualitative inspection of the reconstructed images shows that the $MSOP_{UNET+DEM}$ network produces images better reconstructed than the $MSOP_{UNET}$ and $MSOP_{UNET,w/oSAR}$ networks, especially under thick cloudy areas, highlighting the importance of the SAR and DEM modalities (figure 11).

Model	MSE	SSIM	PSNR	SAM
$SSOP_{UNET,w/oSAR}$	324099	0.8388	24.893	0.1595
$SSOP_{UNET}$	260827	0.8567	25.836	0.1448
$SSOP_{UNET+DEM}$	221909	0.8583	26.538	0.1390
$MSOP_{UNET,w/oSAR}$	141283	0.9249	28.499	0.1128
$MSOP_{UNET}$	138212	0.9267	28.594	0.1111
$MSOP_{UNET+DEM}$	133061	0.9277	28.759	0.1095

Table 8. Comparison between SSOP and MSOP networks, over the MSOP test dataset detailed in section 4.3.2.

5.4.1 Comparison of deep-learning based approaches and gap-filling In this setup, optical images acquired at $t-1$ and $t+1$ are completely cloud-free, which enables the use of the gap-filling. Also, the optical image acquired at t is completely covered by clouds or clouds shadows, according to the cloud masks, helping toward a fair comparison between the approach that consume the optical image at t ($MSOP_{UNET}$) and the gap-filling. We perform the comparison of MSOP, SSOP models and the gap-filling using the $MSOP_{old}$ test dataset detailed in table 6, which matches the validity range of all approaches. Table 9 reports the metrics obtained. We can observe that all the metrics are in favor of the MSOP models. Also, in this particular use-case, the gap-filling leads to superior results to the SSOP models. We can notice that all metrics are largely in favor of the $MSOP_{UNET+DEM}$ model, showing the benefit of the multitemporality, the SAR and the DEM modalities. Figure 12

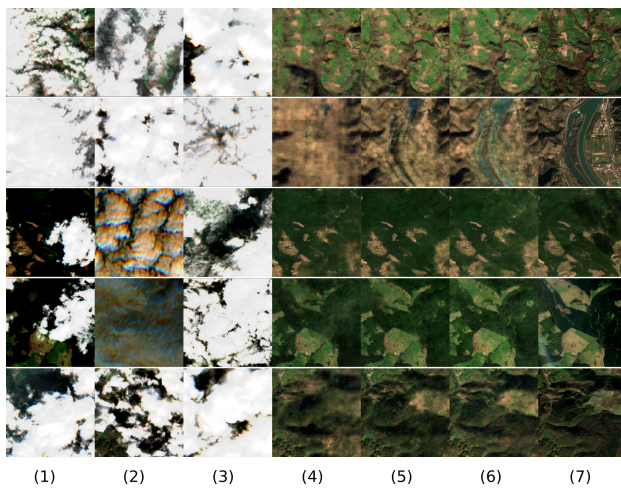


Figure 11. From left to right: input images $S2_{t-1}$ (1), $S2_t$ (2), $S2_{t+1}$ (3), output reconstructed optical images $S2_t$ from $MSOP_{UNET,w/oSAR}$ (4), $MSOP_{UNET}$ (5), $MSOP_{UNET+DEM}$ (6) and the reference image $S2_{t'}$ (7).

shows reconstructed images from the test dataset. We can notice that the gap-filling fails to retrieve various details in the reconstructed images, like sudden crops changes. Also, the input images cloud masks are not always exact, and the gap-filling might interpolates polluted images, unlike the MSOP models which are capable of removing those clouds.

Model	MSE	SSIM	PSNR	SAM
$SSOP_{UNET,w/oSAR}$	239238	0.7911	26.212	0.1847
$SSOP_{UNET}$	178284	0.8187	27.489	0.1472
$SSOP_{UNET+DEM}$	157663	0.8264	28.023	0.1422
gapfilling	79904	0.9249	30.974	0.1021
$MSOP_{UNET,w/oSAR}$	63097	0.9338	32.000	0.0952
$MSOP_{UNET}$	61016	0.9345	32.146	0.0940
$MSOP_{UNET+DEM}$	52814	0.9421	32.772	0.0901

Table 9. Comparison of the gap-filling and the SSOP and MSOP networks. All approaches are compared on the $MSOP_{cld}$ test dataset detailed in table 6.

6. DISCUSSION

We have compared various single date SAR/optical networks, with an ablation study to analyze the contribution of the SAR, optical, and DEM inputs. We have modified the original network from (Meraner et al., 2020), which is considerably greedy in term of computational resources, replacing the ResNet backbone with a U-Net backbone. This has two advantages: first, it is less computationally extensive since convolutions are performed on downsampled features maps. The processing time is diminished with a factor greater than 30, and leads to similar image quality metrics, with a slightly higher PSNR, but slightly lower SAM and SSIM. Secondly, using a U-Net backbone instead of a ResNet backbone enables input images at lower resolution than the 10m bands of Sentinel images: we have shown that a 20m spacing DEM can be injected after the first down-sampling of the network without prior spatial re-sampling, improving the reconstruction of optical images. However, we only have trained all single date based networks using the only the

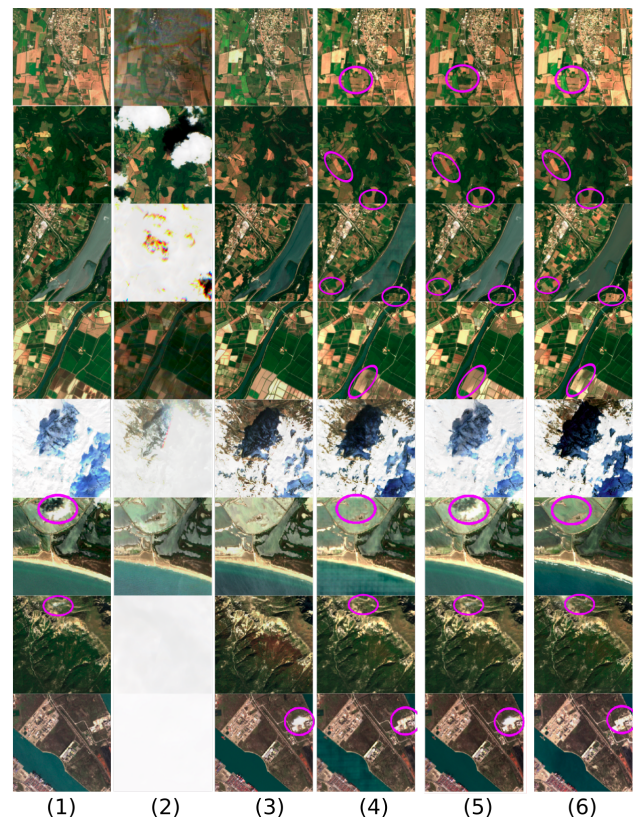


Figure 12. From left to right: input images $S2_{t-1}$ (1), $S2_t$ (2), $S2_{t+1}$ (3), output reconstructed optical images $S2_t$ from $MSOP_{UNET+DEM}$ (4), the Gap-filling (5), and the reference image $S2_{t'}$ (6). In violet are circled details in the reconstructed images that the gap-filling fails to retrieve, or artifacts caused by wrong cloud masks in input images.

l_1 loss, and future works could investigate other objective formulations. We have carried out the comparison of single date networks and the multitemporal networks over the $MSOP$ test dataset, which represents the nominal operational context of both networks, e.g. using every available input images, cloudy or not. Our results shown that the multitemporal networks lead to superior image reconstruction. We believe that more available input images improves the retrieval of the missing contents of the cloudy optical image at t . The comparison between the deep learning based networks and the gap-filling is performed over the $MSOP_{cld}$ test dataset, which contains samples where the t optical image patches are covered by clouds at 100%. The gap-filling performs better than the single date network with a significant margin in this particular setup. However, even though the multitemporal network is not primarily designed for this task, it has outperformed the gap-filling. Finally our ablation study shows that the SAR and the SAR+DEM contribute in the optical image reconstruction process in both single date based networks, and multitemporal based networks. For future works, we believe that a further investigation of the SAR signal contribution should be carried out. For instance, it could be interesting to study if feeding geometrical information (e.g. local SAR incidence angle) in networks would help, and if physical based SAR pre-processing (i.e despeckeling or target decomposition) benefit the optical image reconstruction task.

7. SUMMARY AND CONCLUSION

In this paper, we sought to provide a comparison of single date based and multitemporal convolutional networks with the traditional deterministic temporal interpolation between two images. We have introduced a framework to generate various datasets to train and evaluate various methods for cloudy optical image reconstruction. Our simple yet convenient method relies on space partitioning data structures indexing the crucial parameters of the remote sensing acquisitions, i.e. how SAR and optical remote sensing images must be acquired in the datasets, in term of cloud coverage, SAR/optical maximum gap, number and type of acquisition, and relative acquisition time. We have built several datasets to train single date based networks and multitemporal networks, and to evaluate the different selected approaches, representing various operational contexts. The studied single date based network take their roots in an existing architecture that uses a ResNet backbone, and we have shown how it could be improved using a U-Net backbone, increasing its training and inference speed and enabling to input additional image of different scale. We have built a multitemporal network that generalize the single date image reconstruction from three input pairs of images, and which uses the same backbone shared across the inputs. Our model inputs three cloudy optical and SAR images pairs acquired at dates $t-1$, t and $t+1$, and a DEM. We have lead the comparison between the single date networks, the multitemporal networks, and the gapfilling in various contexts, showing that the gapfilling performs better than the single date based networks in the context of Sentinel-2 time series. We have analyzed the contribution of the different kind of inputs, namely optical, SAR and DEM with an ablation study, showing how the reconstructed image benefits from these modalities. Also, we have shown that, even if the primary design of the multitemporal convolutional network is not focused on image interpolation in temporal domain, it leads to similar even better results than the gap-filling. However, we should interpret our results carefully regarding the ancillary data available for cloud coverage characterization, since our cloud coverage information per patch depends from it, and the bias it can introduce. Finally, we lead our study over a small area that do not represents the various atmospheric conditions all over the earth. With the continuous stream of synchronized SAR and optical acquisitions thanks to the Sentinel constellation, it is expected that future data driven multitemporal models will help to deliver more exploitable data. Our dataset generation framework and our models are available as open-source software³

ACKNOWLEDGEMENTS

The authors would like to thank the reviewers for their valuable suggestions. This work was granted access to the HPC resources of IDRIS under the allocation AD011011608 made by GENCI.

REFERENCES

Baetens, L., Desjardins, C., Hagolle, O., 2019. Validation of copernicus Sentinel-2 cloud masks obtained from MAJA, Sen2Cor, and FMask processors using reference cloud masks generated with a supervised active learning procedure. *Remote Sensing*, 11(4), 433.

Cheng, Q., Shen, H., Zhang, L., Yuan, Q., Zeng, C., 2014. Cloud removal for remotely sensed images by similar pixel replacement guided with a spatio-temporal MRF model. *ISPRS journal of photogrammetry and remote sensing*, 92, 54–68.

Cresson, R., 2018. A framework for remote sensing images processing using deep learning techniques. *IEEE Geoscience and Remote Sensing Letters*, 16(1), 25–29.

Cresson, R., Ienco, D., Gaetano, R., Ose, K., Minh, D. H. T., 2019. Optical image gap filling using deep convolutional autoencoder from optical and radar images. *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, IEEE, 218–221.

Ebel, P., Meraner, A., Schmitt, M., Zhu, X. X., 2020. Multi-sensor data fusion for cloud removal in global and all-season sentinel-2 imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 59(7), 5866–5878.

Eckardt, R., Berger, C., Thiel, C., Schmulius, C., 2013. Removal of optically thick clouds from multi-spectral satellite images using multi-frequency SAR data. *Remote Sensing*, 5(6), 2973–3006.

Farr, T. G., Kobrick, M., 2000. Shuttle Radar Topography Mission produces a wealth of data. *Eos, Transactions American Geophysical Union*, 81(48), 583–585.

Hagolle, O., Sylvander, S., Huc, M., Claverie, M., Clesse, D., Dechoz, C., Lonjou, V., Poulain, V., 2015. SPOT-4 (Take 5): simulation of Sentinel-2 time series on 45 large sites. *Remote sensing*, 7(9), 12242–12264.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hu, G., Li, X., Liang, D., 2015. Thin cloud removal from remote sensing images using multidirectional dual tree complex wavelet transform and transfer least square support vector regression. *Journal of Applied Remote Sensing*, 9(1), 095053.

Inglada, J., Arias, M., Tardy, B., Hagolle, O., Valero, S., Morin, D., Dedieu, G., Sepulcre, G., Bontemps, S., Defourny, P., Koetz, B., 2015. Assessment of an Operational System for Crop Type Map Production Using High Temporal and Spatial Resolution Satellite Optical Imagery. *Remote Sensing*, 7(9), 12356–12379.

Kingma, D. P., Ba, J., 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kruse, F. A., Lefkoff, A., Boardman, J., Heidebrecht, K., Shapiro, A., Barloon, P., Goetz, A., 1993. The spectral image processing system (sips)-interactive visualization and analysis of imaging spectrometer data. *AIP Conference Proceedings*, 283number 1, American Institute of Physics, 192–201.

Li, X., Shen, H., Zhang, L., Zhang, H., Yuan, Q., Yang, G., 2014. Recovering quantitative remote sensing products contaminated by thick clouds and shadows using multitemporal dictionary learning. *IEEE Transactions on Geoscience and Remote Sensing*, 52(11), 7086–7098.

Liu, X., Deng, C., Chanussot, J., Hong, D., Zhao, B., 2019. StfNet: A two-stream convolutional neural network for spatiotemporal image fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 57(9), 6552–6564.

³ <http://github.com/cnes/decloud>

Meraner, A., Ebel, P., Zhu, X. X., Schmitt, M., 2020. Cloud removal in Sentinel-2 imagery using a deep residual neural network and SAR-optical data fusion. *ISPRS Journal of Photogrammetry and Remote Sensing*, 166, 333–346.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical image computing and computer-assisted intervention*, Springer, 234–241.

Sarukkai, V., Jain, A., UzKent, B., Ermon, S., 2020. Cloud removal from satellite images using spatiotemporal generator networks. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1796–1805.

Scarpa, G., Gargiulo, M., Mazza, A., Gaetano, R., 2018. A CNN-Based Fusion Method for Feature Extraction from Sentinel Data. *Remote Sensing*, 10(2), 236.

Shen, H., Li, X., Cheng, Q., Zeng, C., Yang, G., Li, H., Zhang, L., 2015. Missing information reconstruction of remote sensing data: A technical review. *IEEE Geoscience and Remote Sensing Magazine*, 3(3), 61–85.

Wang, Z., Bovik, A. C., Sheikh, H. R., Simoncelli, E. P., 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4), 600–612.

Zhang, Q., Yuan, Q., Zeng, C., Li, X., Wei, Y., 2018. Missing data reconstruction in remote sensing image with a unified spatial–temporal–spectral deep convolutional neural network. *IEEE Transactions on Geoscience and Remote Sensing*, 56(8), 4274–4288.