remote sensing

MDPI

Article

Deep Internal Learning for Inpainting of Cloud-Affected Regions in Satellite Imagery

Mikolaj Czerkawski [1,*], Priti Upadhyay [1], Christopher Davison [1], Astrid Werkmeister [1], Javier Cardona [1,2], Robert Atkinson [1], Craig Michie [1], Ivan Andonovic [1], Malcolm Macdonald [1] and Christos Tachtatzis [1]

[1] Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow G1 1XW, UK; priti.upadhyay@strath.ac.uk (P.U.); christopher.davison@strath.ac.uk (C.D.); astrid.werkmeister@strath.ac.uk (A.W.); j.cardona-amengual@strath.ac.uk (J.C.); robert.atkinson@strath.ac.uk (R.A.); c.michie@strath.ac.uk (C.M.); i.andonovic@strath.ac.uk (I.A.); malcolm.macdonald.102@strath.ac.uk (M.M.); christos.tachtatzis@strath.ac.uk (C.T.)

[2] Department of Chemical and Process Engineering, University of Strathclyde, Glasgow G1 1XJ, UK

* Correspondence: mikolaj.czerkawski@strath.ac.uk

Abstract: Cloud cover remains a significant limitation to a broad range of applications relying on optical remote sensing imagery, including crop identification/yield prediction, climate monitoring, and land cover classification. A common approach to cloud removal treats the problem as an inpainting task and imputes optical data in the cloud-affected regions employing either mosaicing historical data or making use of sensing modalities not impacted by cloud obstructions, such as SAR. Recently, deep learning approaches have been explored in these applications; however, the majority of reported solutions rely on external learning practices, i.e., models trained on fixed datasets. Although these models perform well within the context of a particular dataset, a significant risk of spatial and temporal overfitting exists when applied in different locations or at different times. Here, cloud removal was implemented within an internal learning regime through an inpainting technique based on the deep image prior. The approach was evaluated on both a synthetic dataset with an exact ground truth, as well as real samples. The ability to inpaint the cloud-affected regions for varying weather conditions across a whole year with no prior training was demonstrated, and the performance of the approach was characterised.

Keywords: cloud removal; Sentinel-1; Sentinel-2; deep image prior; internal learning; image inpainting

Citation: Czerkawski, M.; Upadhyay, P.; Davison, C.; Werkmeister, A.; Cardona, J.; Atkinson, R.; Michie, C.; Andonovic, I.; Macdonald, M.; Tachtatzis, C. Deep Internal Learning for Inpainting of Cloud-Affected Regions in Satellite Imagery. *Remote Sens.* **2022**, *14*, 1342. https://doi.org/10.3390/rs14061342

Academic Editor: Akram Al-Hourani

Received: 14 February 2022
Accepted: 8 March 2022
Published: 10 March 2022
Corrected: 14 December 2022

1. Introduction

Cloud occlusion reduces the temporal availability and, in turn, the usability of optical satellite data, imposing a significant obstacle to applications where frequent sampling of the ground surface is necessary. In the field of precision agriculture, for tasks such as monitoring crop growth [1], crop classification [2], or crop yield prediction [3], gaps in data result in challenges related to the development of accurate models, variability in prediction performance, and the need for the use of data imputation. Open data sources, such as those generated by the EU Copernicus Sentinel missions [4], have been instrumental to the accelerated development of applications, enabling easy and wide access to large sets of high-resolution data at relatively short revisit periods. However, the fundamental limitations imposed by cloud obstruction remain, which can be mitigated either through the development of new models that do not rely exclusively on optical data or by techniques that reconstruct optical imagery for cloud-affected regions.

Reported approaches to cloud removal from optical satellite images can be segmented into a number of strands. Most early developments, before the growth of deep-learning-based approaches, used interpolation-based techniques [5–7], filtering [8,9], or the creation of image mosaics or composites [10–15]. A number of approaches employing ma-

chine learning were reported [16–19] prior to a notable increase in research to determine the potential of deep neural networks to yield advances in performance [20–30]. Although a subset of the proposed methods attempt to remove clouds using a single image [8,20,21,31–33], more common approaches either rely on multi-temporal data to inform the reconstruction [5,7,9–19,25,28,29,34–39] or use sensing modalities not impacted by clouds [6,22,23,25–27,30,40].

Most deep-learning-based cloud removal methods depend on a form of external training [20–30,41,42], relying on a pre-training phase, during which the network is optimised to produce a desired target output for a fixed dataset. As a consequence, the performance and generalisation of the models depend heavily on the level of representation the fixed dataset provides of the conditions encountered in deployment. The training dataset must be sufficiently diverse to capture the complexity of the real-world application so as to minimise the expectation of loss within the operational scenario. Conversely, the consequence of striving to achieve increased diversity is a trade in performance on individual samples in favour of average accuracy.

Several deep learning approaches [27,30] claim that larger datasets result in neural network models able to learn general and adaptable transformations. However, this assertion is at odds with the fact that in practice, the majority of datasets contain bias and a considerable amount of features that are highly predictive within the dataset, but do not generalise to new samples [43]. Thus, methods that do not rely on a process of optimisation on a fixed number of samples [5–15], but use engineered priors that guide the synthesis process are more attractive. Recently, an approach [29] that combines external learning with prior-based synthesis has been proposed where spatio-temporal representations are reconstructed with a Deep Image Prior (DIP), while an externally trained network enforces the context consistency of the synthesised images. It must be noted, however, that the external network may introduce a bias in a similar fashion to other external learning approaches.

Here, an alternative technique embracing a pure internal learning paradigm that relies on the optimisation of the network on a single sample of interest and thus obviates the need for pre-training on an external dataset is detailed and evaluated. The approach is restricted to model tasks where a significant part of the information required for reconstruction is not obscured in the source sample. Within the satellite imaging context, data are available from multi-temporal and multi-source streams, enabling the application of knowledge arising from solutions to a number of well-known computer vision inverse problems [44–47] such as image completion, super-resolution, and denoising. To date, only one other reported technique has adopted the proposed methodology [48], employing a sequence-to-sequence internal learning scheme to translate sequences of Synthetic Aperture Radar (SAR) satellite images to sequences of multi-spectral images using a 3D U-Net architecture. However, the approach of [48] is fundamentally different from deep image prior work [44] and the framework proposed here as it feeds the informing SAR signal directly to the reconstruction network instead of employing noise tensor input and reconstructing the informing signal along with the target optical representation. It is not demonstrated why this should be beneficial in the internal learning context. The sequence-to-sequence model does, however, limit the flexibility of the solution as the data are constrained to the same number of SAR and optical samples. Furthermore, the method was evaluated only on three sample sequences. During the review stage of this manuscript, an independent study on the application of the deep image prior has been published [49]. The paper reported on multi-temporal reconstruction and relied on a recent cloud-free reference, evaluated on 10 Landsat-8 and 10 Sentinel-2 patches. Here, the use of the multi-temporal historical average signal or multi-source SAR information is proposed, which avoids issues with overlapping cloud regions. Furthermore, this work performed a sensitivity analysis of the cloud mask size and the synthesis quality.

In this paper, an end-to-end workflow for the removal of cloud cover in Sentinel-2 optical satellite images is presented for precision agriculture applications, where flexibility and adaptation to a target fixed region of interest is a key requirement. Several methods of informing the reconstruction process were tested, including both multi-source and multi-temporal approaches; for the former case, SAR images from Sentinel-1 were used as an informing input.

The proposed solution is an application of the DIP methodology [44], a prominent internal deep learning approach for image synthesis. Furthermore, the contribution of this work involved the analysis and comparison of several processing approaches to the problem, depending on what data sources were used. Evidence is provided that the proposed internal techniques are capable of producing high-quality reconstructions, which was tested on a synthetic dataset with an accessible ground truth, as well as on real images. Finally, unlike other deep learning methods for cloud removal, the solution is readily adaptable with no additional training, be it using top-of-atmosphere or bottom-of-atmosphere hyperspectral or, indeed, a number of other sensor data and irrespective of the preprocessing approach employed.

The Materials and Methods Section (Section 2) provides details on the data pre-processing routine and the creation of the synthetic dataset, followed by a description of the proposed architectures (Section 2.2). Section 3 presents both quantitative results (Section 3.1), where the performance metrics are computed for the entire dataset, as well as qualitative results (Section 3.2). Conclusions are drawn in Section 4.

2. Materials and Methods

In this section, the dataset used for the evaluation experiments is described, followed by a segment detailing the proposed model architectures.

2.1. Supporting Data

The focus was the removal of clouds from optical Sentinel-2 images in the context of crop monitoring. Hence, the dataset created for the development was based on Sentinel-2 data from a Level-2A product, offering bottom-of-atmosphere reflectance measurements, proven to be most appropriate in vegetation monitoring applications as opposed to top-of-atmosphere data. Even so, the internal synthesis approach presented is adaptable as it is based on the content of the target image only and, as a consequence, can be directly applied to other image types, such as the more commonly used Level-1C Sentinel-2 images. In fact, results (Section 3.2) are also reported on the performance of the method with Level-1C samples for comparison with an external model. Finally, the resultant Sentinel-2 images were paired with Sentinel-1 equivalents based on the closest temporal proximity.

The model requires a cloud mask (or cloud and shadow mask), and although a significant number of cloud detection algorithms have been reported to date [50], the goal was to evaluate the performance of the proposed method on the inpainting task alone. To avoid the propagation of errors from the cloud detection models, the inpainting regions were obtained by removing portions of images under clear sky, providing an accessible and precise ground truth.

The dataset comprises two-year temporal coverage measurements for two different regions, one in Scotland, the other in India, each containing approximately 200 samples (data available at https://doi.org/10.5281/zenodo.5897695 (accessed on 9 February 2022)). The extent of the data allows the valid test of the approach over a substantial period of time subject to varying seasonal conditions and across two disparate geographical locations. Every image is effectively a test sample, obviating the need for splitting the data into training and test subsets, since an internal learning approach was employed. All available images for the years 2019 and 2020 were manually assessed to extract a subset of clear sky images to ensure that the samples were clouds-free. All clear sky images from 2019 were used as a source of prior information, and the images from year 2020 were used for the evaluation. Cloud regions were simulated by obtaining plausible cloud masks from

Sentinel-2 data of the Scotland region with a coverage area between 10% and 50%, resulting in 17 realistic mask shapes. The combination of clear sky images with each cloud mask yielded the evaluation dataset. The resulting sample counts for both regions are shown in Table 1.

Table 1. Resulting sample counts for each dataset. The bottom row contains the total count for 17 cloud masks.

Dataset	Scotland	India
Informing Samples (2019)	18	34
Inference Samples (2020)	20	30
Total Task Samples	340	510

The raw data are acquired and preprocessed using the analysis-ready data framework described in [51], as shown in Figure 1. The Sentinel-1 SAR data were radiometrically calibrated, speckle filtered, and terrain corrected. The Sentinel-2 multi-spectral bands and cloud masks (opaque cloud and cirrus) from the Level-2A products were re-sampled to a 10 m spatial resolution. Sentinel-1 and Sentinel-2 tiles were then co-located, yielding image pairs of 256 × 256 px. The output pairs were then treated in a manner similar to [27]. Sentinel-1 images were (1) converted to a decibel scale, (2) clipped (VV range: $[-25.0, 0]$, VH range: $[-32.5, 0]$), and shifted to $[0, 2]$. Sentinel-2 images were clipped between $[0, 10^4]$ for all optical channels and divided by 2000.

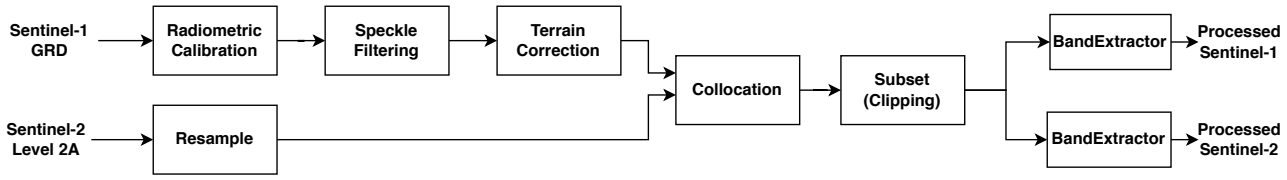


Figure 1. Snap processing pipeline for Sentinel-1 Ground Range Detected (GRD) and Sentinel-2 Level-2A products.

2.2. Architecture

The internal cloud removal architectures proposed here rely on the techniques proposed in [44], broadly referred to as Deep Image Prior (DIP). The key premise of the approach is that signals, especially natural images, can be encoded in the weights of a deep convolutional neural network by overfitting the model to generate the signal of interest for a given constant input. The network weights tend to converge to solutions of lower entropy [44], where all textures in the image are derived from a relatively narrow distribution, rendering the DIP method applicable to several image inverse problems, such as image inpainting, denoising, or super-resolution. DIP enables a significant amount of prior knowledge on natural images to be introduced through the use of a convolutional architecture.

The applications demonstrated in the original DIP work operate on a single image; here, the method was extended to cloud removal by using additional information, viz. involving several images. The additional images can be obtained from a different sensor (multi-source synthesis) or from different points in time (multi-temporal synthesis). The results demonstrated that the DIP approach can be effectively applied to both multi-temporal and multi-source satellite images through the stacking of all frames to create an image representation with more channels. Furthermore, good-quality images were produced even when a disparity between domains occurred, as is the case for optical and SAR sensor data.

A number of DIP-based reconstruction methods have been proposed, depicted in Figure 2. The most direct approach applies the DIP network to the masked hyperspectral image alone, as shown in Figure 2a, in alignment with the default motivation of the technique, where the inpainting is based only on the textures present in the non-masked region,

i.e., not obscured by clouds. Two approaches can be used to incorporate an additional data source: either through the optimisation of (i) a residual target output (Figure 2b,c) or (ii) a stack of the target image with the supporting information appended as additional channels (Figure 2d–f). For the residual modes, the network is tasked with computing the difference between either a sample from a different source (to achieve the same number of channels (3) necessary for the multi-source residual, an additional channel containing the VV and VH mean was appended to the SAR representation) (Figure 2b) or a sample from a different time (Figure 2c). The residual component image that the core DIP network was optimised to produce can be expected to exhibit a similarly narrow texture distribution, just as the reconstructed Sentinel-2 image. For the stacked modes (Figure 2d–f), the supporting information could be multi-source (d), multi-temporal (e), or a combination thereof (f). For these cases, the mask was only applied to the cloudy query image so the supporting data from the synthesised region could be backpropagated.

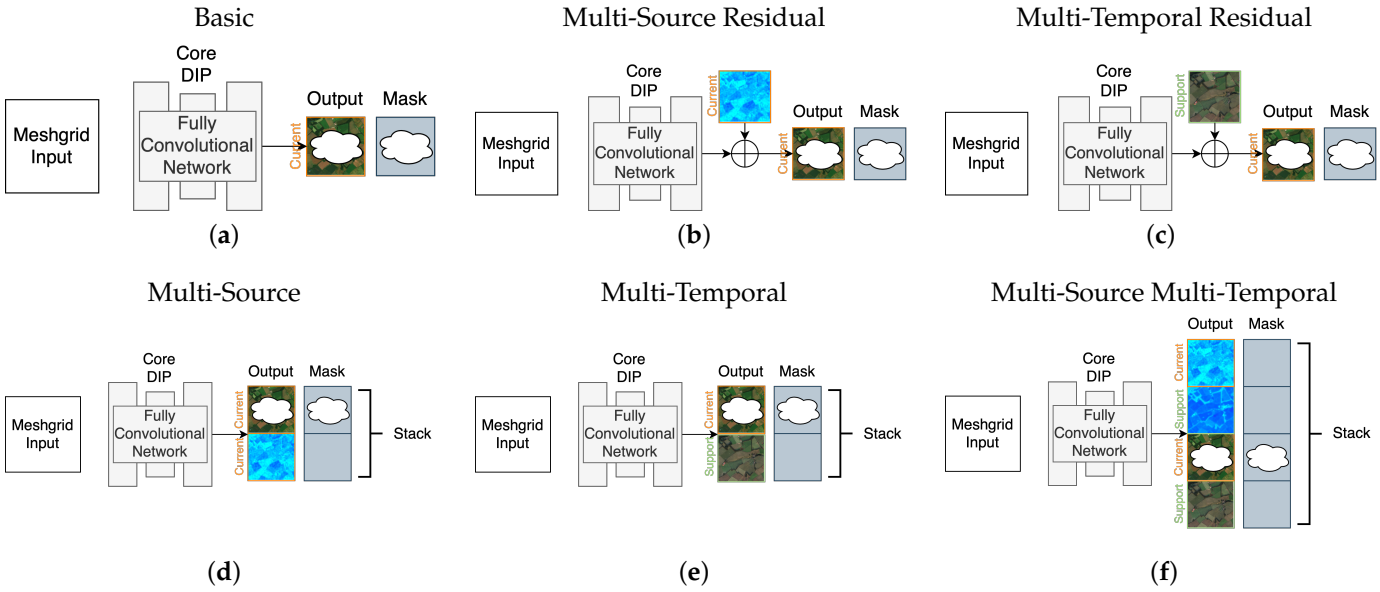


Figure 2. Diagrams of synthesis modes: (**a**) single image, (**b**) multi-source residual, (**c**) multi-temporal residual, (**d**) stacked multi-source, (**e**) stacked multi-temporal, and (**f**) stacked multi-source multi-temporal.

Multi-temporal synthesis requires additional temporal samples to be used at the inference time. This additional source can be static (the same reference used for a number of different predictions) or dynamic, where an appropriate reference is chosen specifically for the synthesised sample (for example, the last clear sky image taken before the cloudy sample). The conducted experiments involved the former, to eliminate the deviation in the reference and allow a consistent architectural assessment. In order to create a static reference, all clear sky images from the year 2019 were averaged. The main purpose of taking a mean over the entire year comprising multiple seasons and weather conditions was to extract the static structural component of a given region. Moreover, taking the mean also had an additional denoising effect, especially for SAR images. The resulting mean representations of the informing samples are shown in Figure 3.

Cloud removal for Sentinel-2 data can be achieved by utilising Sentinel-1 SAR data as the alternative sensing source. The SAR component was incorporated as a 2-channel image (VV, VH) to enable a valid comparison with other cloud removal techniques [27]. Several approaches are possible to generate temporal support data. A source of dynamic time-dependent support can be obtained from the last Sentinel-2 observation; however, there is no guarantee that this image will not contain significant cloud cover. Thus, the last cloud-free (or nearly cloud-free) observation could be used, but that may lead to temporal inconsistencies between the inferred samples and their support counterparts, compromising

performance stability. An alternative is to use a mean support sample derived as the average signal from the cloud-free images captured over the preceding year of the inference period. The average Sentinel-1 and Sentinel-2 data will contain an approximation of the static component shared by all images over the year. This way, each inferred sample was biased by the same support signal. Examples of the mean support data are shown in Figure 3.

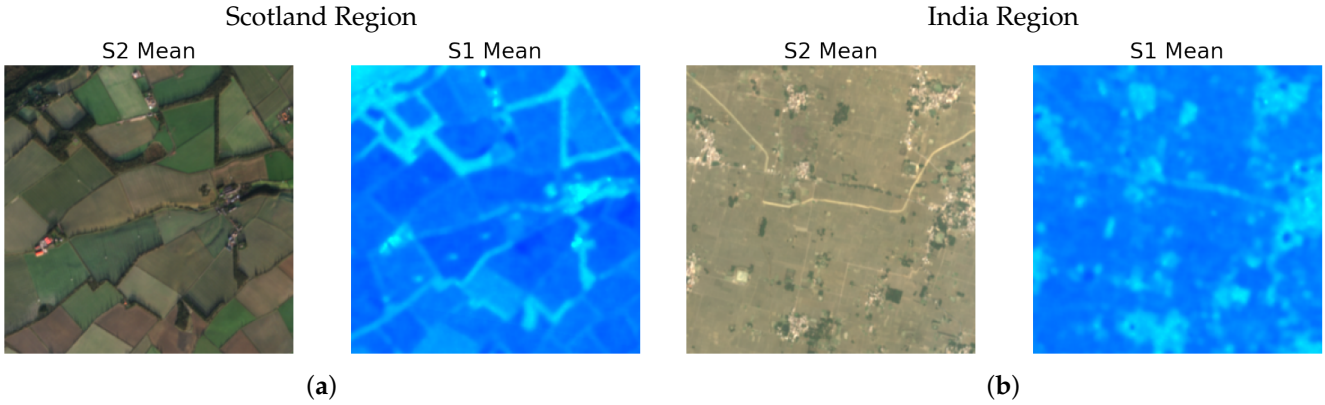


Figure 3. Mean images for year 2019 for the region in Scotland (**a**) and the region in India (**b**). Sentinel-1 VV and VH are displayed as green and blue, respectively.

All the experiments conducted were obtained using an Adam optimiser with exponential decay rates for the first and second moment coefficients $\beta_1 = 0.9$, $\beta_2 = 0.999$, a stabilising term $\epsilon = 10^{-8}$, and a learning rate of 2×10^{-2}. The networks were trained for 4000 steps with a regularisation noise of 10^{-1}. The code for the proposed method is available at https://github.com/cidcom/satellite-cloud-removal-dip (accessed on 9 February 2022). The result of an ablation study is shown in Figure 4, demonstrating the visual quality of the synthesised cloud inpainting for all synthesis modes presented in Figure 2. Two performance metrics were employed for evaluation: Structural Similarity Index (SSIM) [52] as an indicator of structural coherence and Root-Mean-Squared Error (RMSE) as a measure of reflectance error. Tests were carried out by repeating each mode 4 times, and the displayed examples corresponded to the run with the median SSIM, applied to the inpainted region only, unless otherwise stated. The basic scenario with no extra informing data produced results that matched the colour tones of the image, but lacked fine detail. The residual multi-source approach produced poor-quality results and may require additional mechanisms to achieve improved synthesis. The multi-temporal residual mode produced a reasonable quality result, but was constrained to the same number of synthesised channels as the reference representation, unlike the stacked modes, where the output can be adjusted to yield any number of frames and channels. The stacking approaches are presented in the bottom row, leading to a considerable improvement when employing multi-temporal data.

The metrics between two images I_A and I_B (which correspond to a test image and appropriate ground truth image) were computed using the following formulas:

$$\mathrm{RMSE}(I_A, I_B) = \sqrt{\frac{(I_A - I_B)^2}{H \times W}} \tag{1}$$

where H is the height and W is the width of both I_A and I_B.

$$\mathrm{SSIM}(x, y) = \frac{1}{|\Pi(I_A, I_B)|} \sum_{x,y}^{\Pi(I_A, I_B)} \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{2}$$

where x, y correspond to individual patches of size 11×11 in I_A and I_B (with Gaussian window as described in [52]), respectively, where $\Pi(I_A, I_B)$ is the set of all collocated patches and $|\Pi(I_A, I_B)|$ the total number of elements in that set. The variables with μ correspond to the mean value, σ to the variance, and σ_{xy} to the covariance. Standard constant values of c_1 and c_2 were used for computational stability, as proposed in [52].

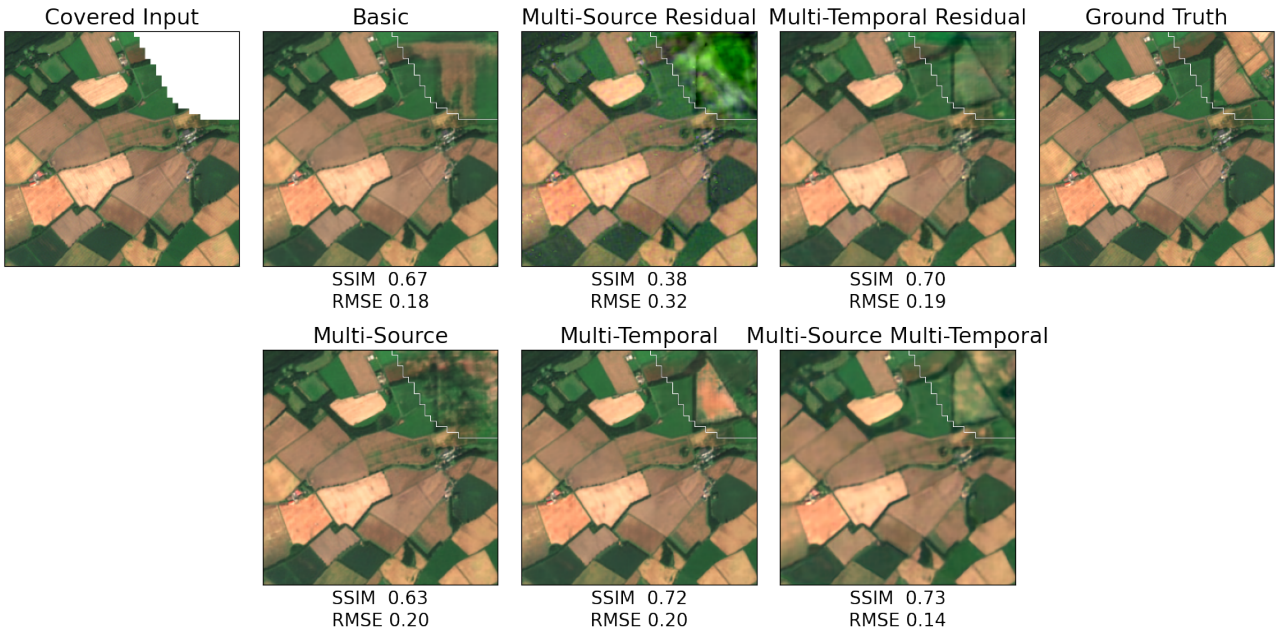


Figure 4. Comparison of DIP-based synthesis approaches. Median score samples from 4 repeated runs are shown. Metric values are given for the inpainting region.

3. Results

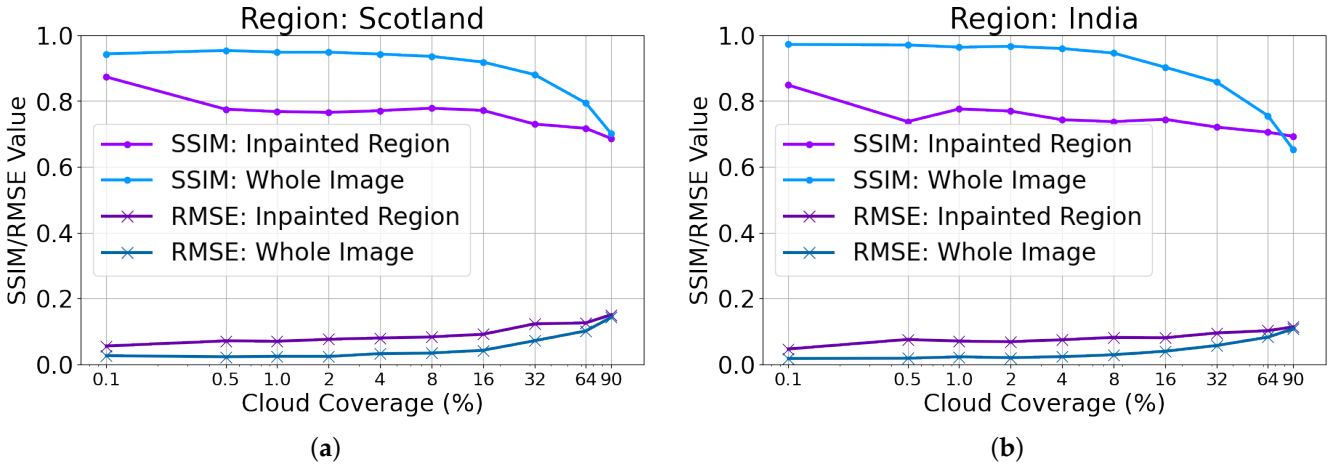
The quantitative results relied on the dataset described in Section 2, where clear sky samples were reconstructed from artificially masked images. The qualitative results are presented to demonstrate the visual appearance of the synthesised regions.

3.1. Quantitative Results

All synthesis modes were applied to the evaluation dataset containing the two regions of Scotland and India, with performance metrics computed for the image as a whole and for the inpainted region only; the results of the inference are shown in Table 2. The stacked MT and MS-MT variants that utilised multi-temporal data consistently outperformed the other variants for both datasets and for both the full image and inpainted regions (A metric applied to the inpainted region measures the quality of new generated data and, when applied to the whole image, quantifies the distortion introduced by the optimisation process. The distortion can be partially prevented by stitching the uncovered region with the synthesised pixels, but new distortions on the stitch borders may be introduced by the process). This suggests that the inclusion of the multi-temporal component is more beneficial for the considered synthesis framework. Both residual modes exhibited inferior performance compared to the corresponding stacked modes that used equivalent data. Furthermore, the residual modes constrained the number of frames and channels, and thus, the remainder of the evaluation was focused on the stacked synthesis modes.

The evaluation dataset was designed to contain real cloud masks with coverage from a selected range (10–50%), useful for simulating real-world conditions, but does not give an opportunity to validate how the system performance varies depending on the size of an inpainted region. Consequently, another experiment was carried out, where clear sky samples from each of the two datasets were reconstructed for a number of artificial cloud masks with a controlled coverage area. The resulting SSIM and RMSE traces are shown in

Figure 5. The inpainting quality appears to be consistent for cloud coverage ratio between 0.1% to 16%, where the inpainted region SSIM is approximately 0.75 for the Scotland and India regions. The SSIM decreased beyond 16% cloud coverage, however, but never fell below 0.6. Naturally, the whole-image SSIM was higher than that of the inpainted region since the whole image contained the noncovered region, which was optimised directly. The visual quality of the reconstruction of selected samples from each dataset for a range of cloud coverage levels is illustrated in Figure 6.



(**a**) (**b**)

Figure 5. SSIM and RMSE plots for the cloud coverage sweeps for MS-MT mode for the (**a**) Scotland region and (**b**) India region.

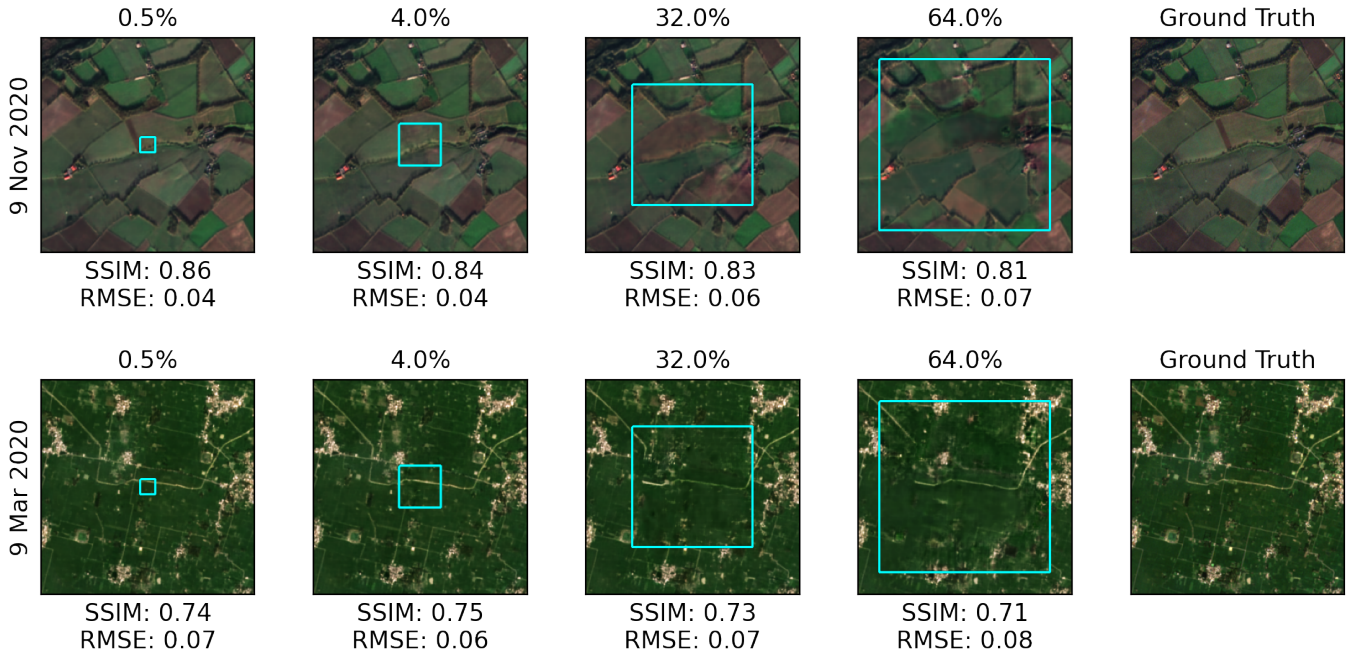


Figure 6. Samples of images reconstructed in the cloud coverage sweep procedure (ranging from 0.5% to 64.0% cloud cover) for MS-MT mode in Scotland (**top row**) and India (**bottom row**). The SSIM and RMSE metric values are shown for the inpainted region.

Table 2. MS—Multi-Source (Sentinel 1), MT—Multi-Temporal, -R—Residual Variants. Optimal performance maximizes SSIM and minimizes RMSE. This occurs for the modes using temporal support data, with the MS-MT mode surpassing all others for both datasets and both for the inpainting region and the whole image.

Dataset			Direct Modes			Stacked Modes		
			Basic	**MS-R**	**MT-R**	**MS**	**MT**	**MS-MT**
Scotland	Whole	SSIM ↑	0.84	0.73	0.86	0.84	0.88	0.88
		RMSE ↓	0.09	0.13	0.08	0.08	0.07	0.07
	Inpainting	SSIM ↑	0.58	0.33	0.66	0.58	0.73	0.73
		RMSE ↓	0.15	0.23	0.15	0.14	0.12	0.11
India	Whole	SSIM ↑	0.84	0.74	0.84	0.84	0.87	0.87
		RMSE ↓	0.07	0.12	0.08	0.07	0.06	0.05
	Inpainting	SSIM ↑	0.53	0.32	0.57	0.53	0.64	0.66
		RMSE ↓	0.13	0.21	0.15	0.12	0.11	0.09

3.2. Qualitative Results

Figures 7 and 8 present a compilation of the results distributed across the full year for the Scottish and Indian regions. It is clearly evident that the synthesis modes adjusted the content of the inpainted region based on the information available in the clear region, and the informing reference acted as a guide for generating the static structure of the image. The Multi-Source (MS) synthesis mode yielded fairly poor structural coherence, indicated by consistently lower SSIM scores, also apparent upon visual inspection. Both MT and MS-MT modes generated images with much higher structural coherence and achieved a comparable level of performance. The main differences between the two modes can be observed in the colours of individual objects in the inpainted region. The MT mode had no recent (i.e., temporally proximate) reference of the inpainted regions and relied only on historical data and on what was available outside of the cloud mask. On the other hand, the MS-MT mode did entail access to temporally proximate SAR data, but the different sensing modality contributed to textural distortions in the reconstruction of the optical data. Similar textural distortions were visible in both MS and MS-MT modes, but were less pronounced in the latter owing to the inclusion of the temporal source. Finally, some samples were more challenging to synthesise. For instance, the sample of 28 December 2020 in Scotland (bottom row of Figure 7) contained a significant amount of snow cover. The image contained a strong bias towards a value of saturated white with low variance, posing a risk of unstable optimisation, which occurred in the MT mode for this particular sample.

The influence of image content on the reconstruction process was accentuated when the two geographical regions were compared. The Scotland region consisted of considerably larger objects, yielding images that were inherently easier to fit as an output of a Convolutional Neural Network (CNN). This was confirmed by the higher performance achieved for the Scotland region in Table 2. The amount of fine structural detail can have a considerable effect on the robustness of each tested synthesis mode. In this case, the information from the SAR sensor, while useful for the large-scale coherence, contained noise and had insufficient resolution, resulting in higher distortions of the inpainted regions, an effect visible in Figure 8, where the MT approach provided more consistent inpaintings than the other two approaches that used SAR data.

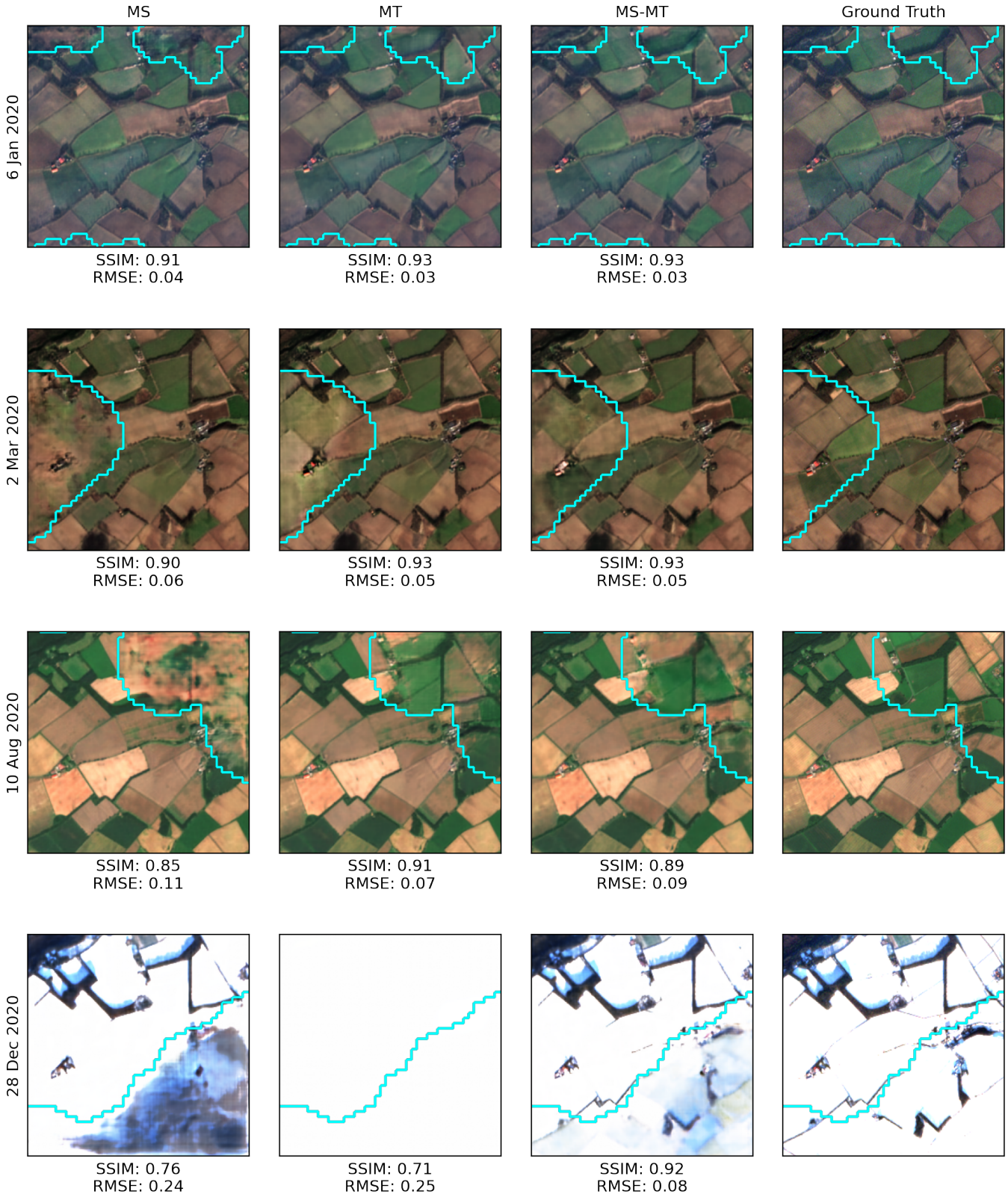


Figure 7. A selection of reconstructed samples across the full year in the Scotland region. This demonstrates the seasonal adaptability of the proposed methods. The cyan line delineates the cloud-affected from the cloud-free regions in the image.

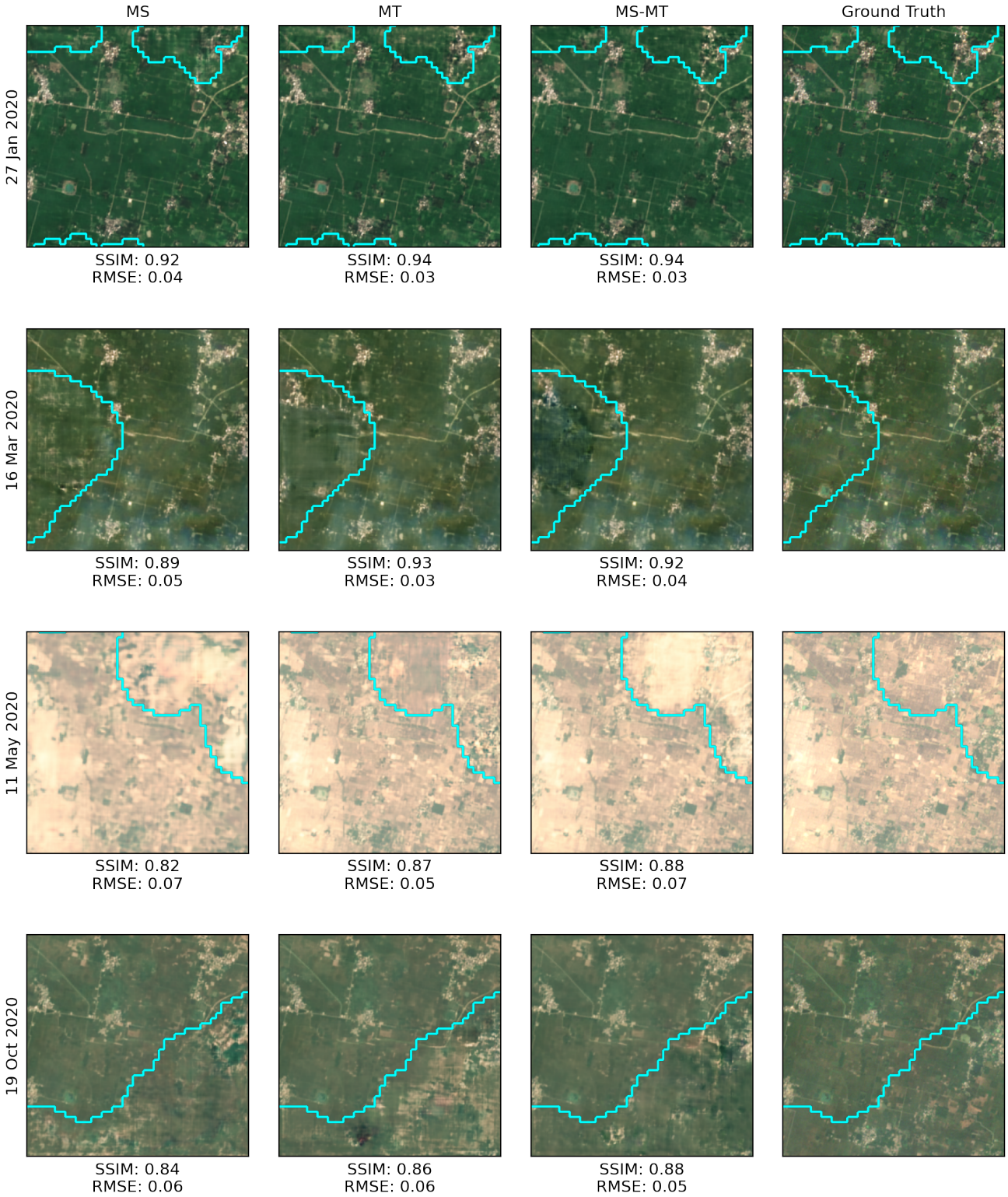


Figure 8. A selection of reconstructed samples across the full year in the India region. This demonstrates the seasonal adaptability of the proposed methods. The cyan line delineates the cloud-affected from the cloud-free regions in the image.

The best- and worst-scored samples for both datasets are shown in Figures 9 and 10 for Scotland and India, respectively. Note that for the Scotland region, the snowy samples from 28 December were excluded when selecting extreme results, as the severe saturation of the images naturally yielded a high SSIM for the reconstructions. In Figure 9, the first

two rows contain the samples yielding the worst performance, while the bottom two rows contain the best. Interestingly, a degree of overlap between the identified extremes for the individual modes can be identified. For instance, all modes share the same date for the maximum and minimum RMSE scores, as well as for the maximum SSIM score in Figure 10, potentially attributable to a combination of factors such as the specific image content, its relation to the informing reference, and the mask area.

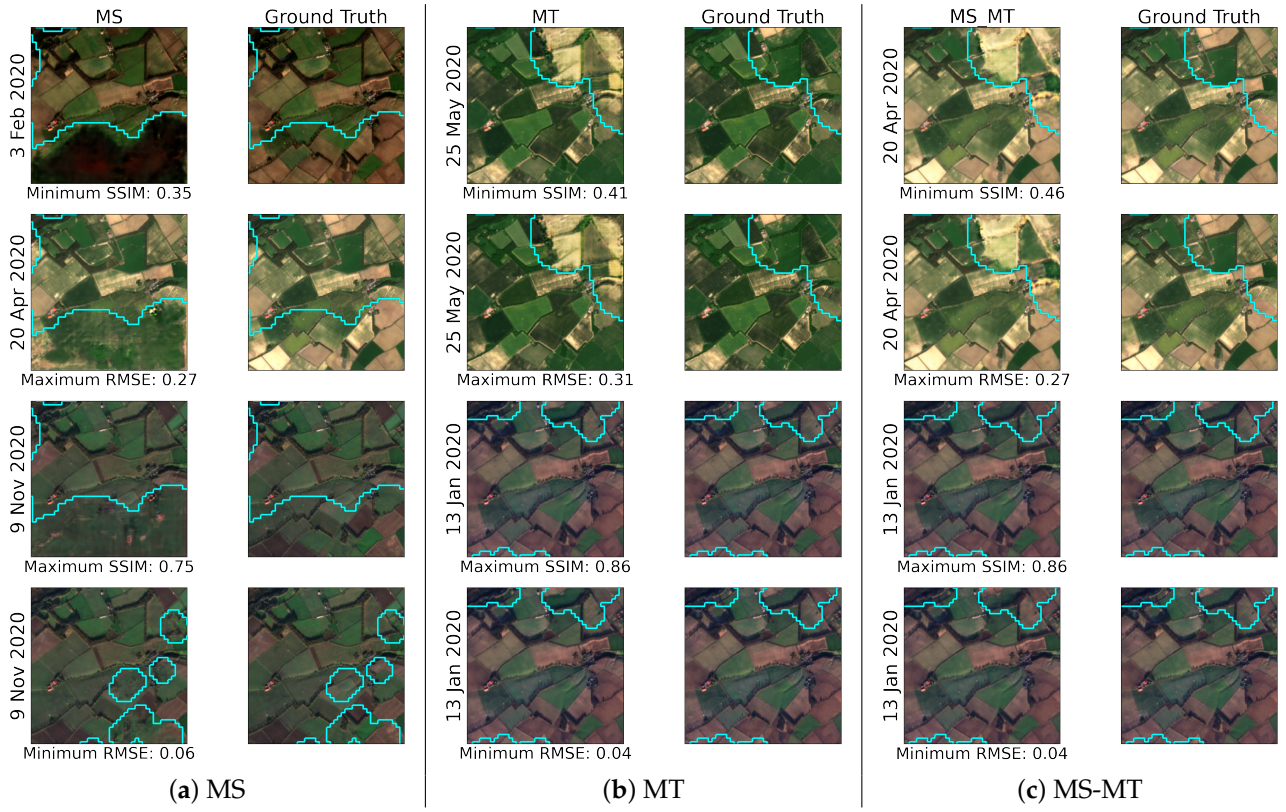


Figure 9. Reconstructions with the lowest (top two rows) and highest (bottom two rows) performance for both the SSIM and RMSE for the Scotland region. Each row contains all three stacked modes (MS, MT, and MS-MT).

3.3. Application in a Blind Setting

The proposed techniques do not require pre-training and can be readily applied to new settings. To demonstrate this capability, the methods were compared to an existing cloud removal model of DSen2-CR [27]. The pre-trained DSen2-CR uses Level-1C Sentinel-2 images containing a top-of-atmosphere view. A small number of Level-1C top-of-atmosphere cloud-affected images were obtained: three for both Scotland and India. Although the DIP-based methods require cloud masks, it is nevertheless possible to use existing cloud detection algorithms; however, these are likely to produce false negatives (i.e., undetected cloud regions). Leakage of cloud components outside the mask can affect the optimisation process of the proposed synthesis methods by distorting the texture distribution of the image. For that reason, the six images used in the evaluation were manually annotated to ensure no leakage.

Figures 11 and 12 contain a comparison of the three considered synthesis modes (multi-source, multi-temporal, and MS-MT) to an externally trained model of DSen2-CR for Scotland and India, respectively. Generally, DSen2-CR generated clear-sky regions that were very close to the S2 input; yet, dark artefacts were evident in the clouded regions that poorly matched the colour distribution. The MS mode using equivalent input data (with the addition of a cloud mask) did not produce the artefacts in the clouded regions. The modes MT and MS-MT provided a more detailed structure in the synthesised output, but relied on additional information.

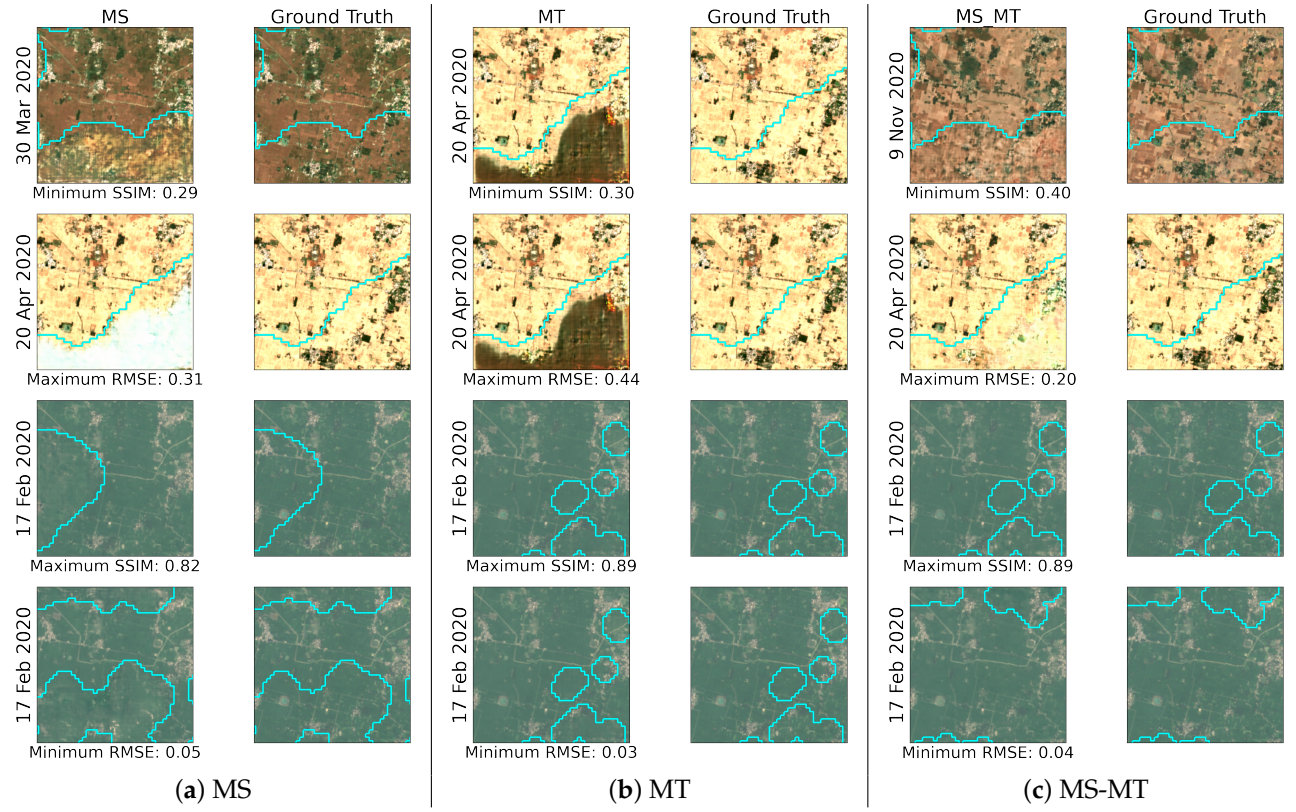


Figure 10. Reconstructions with the lowest (top two rows) and highest (bottom two rows) performance for both the SSIM and RMSE for the India region. Each row contains all three stacked modes (MS, MT, and MS-MT).

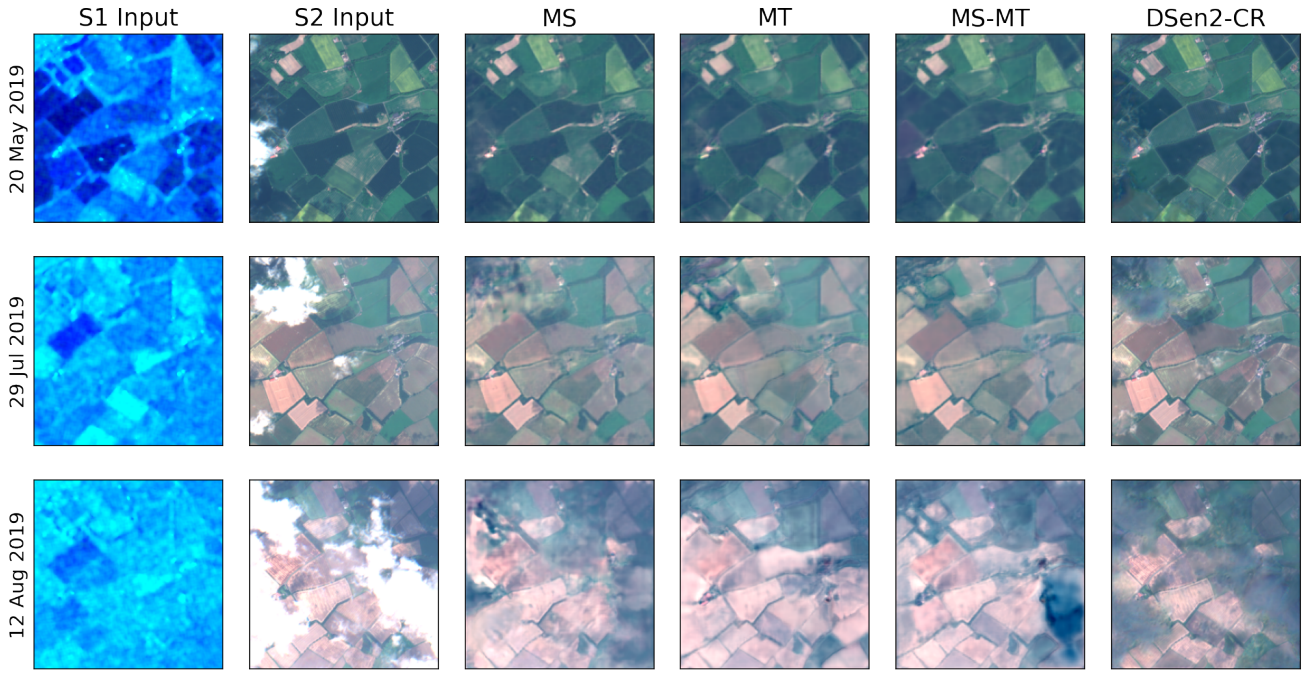


Figure 11. Results on the blind cloud removal task, with no available clear sky references for the Scotland region. The cloud masks were annotated manually to enable a comparison with the DSen2-CR external method.

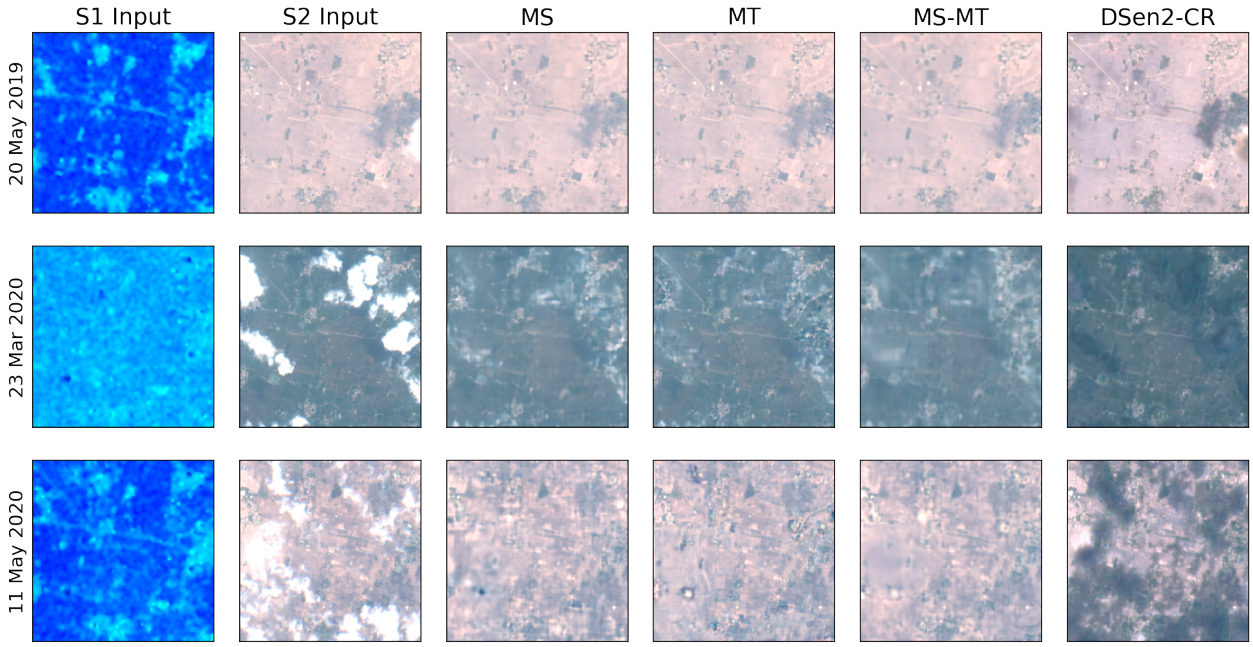


Figure 12. Results on the blind cloud removal task, with no available clear sky references for the India region. The cloud masks were annotated manually to enable the comparison with the DSen2-CR external method.

4. Conclusions

The key factor differentiating the presented internal learning approach to cloud region inpainting in satellite imagery from solutions relying on external learning is that it permits the development of a flexible model not biased by the dataset. Elimination of the cost and time of pre-training, as well as the sourcing, preparation, and curation of the training dataset are an attractive set of advantages. Furthermore, inference takes in the order of two minutes on a consumer-grade GPU (24GB NVIDIA Titan RTX), acceptable for almost all agriculture applications considering the significantly higher delays already present in satellite processing pipelines. For scenarios where processing of larger areas is required, this pipeline can be fully parallelised on a high-performance computing platform to yield a comparable processing time, assuming that an automatic cloud detection module is employed (more details on the processing time are included in Appendix A). The key advantage is that the proposed approach is modality agnostic and can be readily applied to different signal types with no requirement for a new dataset or additional training, particularly beneficial for settings where labelled data are not available.

The solutions presented adopted several approaches to inform the synthesis process. The results showed that the DIP-based methods handle stacked data of the same modality better than multi-source composites. Consequently, historical optical data yielded higher-quality synthesised samples than the two-band SAR reference, potentially attributable to the distortions present in the employed SAR representation. Potential solutions to mitigate this effect are either to devise alternative architectures or improved SAR representation or preprocessing. The advantage of a multi-temporal reference, observed in the context of the presented developments, aligns with the early cloud removal solutions that also employ historical data of the same modality. In this work, the benefit of the temporal reference was maintained even when the time difference between the synthesised image and the reference was as large as 1 y.

The proposed synthesis modes were evaluated both quantitatively and qualitatively for two geographic locations of Scotland and India on data obtained throughout the year 2020. The multi-temporal and multi-source (MS-MT) mode offers the highest performance of the three evaluated modes with an average SSIM greater than 0.87 and 0.64 for the

whole image and inpainting regions, respectively. The Multi-Temporal only mode (MT) offered performance on par with MS-MT. The MS performed significantly worse than the temporal approaches. Furthermore, the MS-MT mode was tested for varying degrees of cloud coverage, exhibiting a stable performance even when more than 16% of pixels were inpainted. All three proposed internal synthesis modes were qualitatively compared to an externally trained network to demonstrate the synthesis quality.

The techniques presented all rely on the application of high-quality cloud masks and thus on a robust cloud detection scheme. Furthermore, the results indicated that the synthesised images can vary depending on the run. While the characteristics of the inpainting task impose a wide range of feasible solutions, further work is required to reduce the distortion and variability of the synthesised outputs.

Author Contributions: Conceptualisation, M.C. and C.T.; methodology, M.C. and P.U.; software, M.C., P.U. and C.D.; validation, M.C.; formal analysis, M.C.; investigation, M.C. and P.U.; data curation, P.U., A.W., M.C. and C.D.; writing—original draft preparation, M.C.; writing—review and editing, M.C., P.U., J.C., M.M., C.T. and I.A.; visualisation, M.C.; supervision, C.T., J.C., C.M. and I.A.; project administration, M.C. and C.T.; funding acquisition, I.A., C.M., R.A. and C.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the European Union Horizon 2020 Research and Innovation Programme, Grant Number 825355.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset comprises two-year temporal coverage measurements for two different regions, one in Scotland, the other in India, each containing approximately 200 samples. Data is available at https://doi.org/10.5281/zenodo.5897695 (accessed on 9 February 2022). The code is openly available at https://github.com/cidcom/satellite-cloud-removal-dip (accessed on 9 February 2022).

Acknowledgments: The work was partially funded by the European Union Horizon 2020 Research and Innovation Programme "Fostering Precision Agriculture and Livestock Farming through Secure Access to Large-Scale HPC-Enabled Virtual Industrial Experimentation Environment Empowering Scalable Big Data Analytics (H2020-ICT-2018-2020) (CYBELE)" under Grant Agreement No. 825355.

Conflicts of Interest: The authors declare no conflict of interest. The funder had no role in the design of the study; in the collection, analyses, or interpretation of the data; in the writing of the manuscript; nor in the decision to publish the results.

Appendix A. Technique Scalability

This section provides details on the time and memory cost of the current method without any attempt to optimise the network for large-scale data processing and comments on the result. Indicative results for a region of 100 × 100 km^2 (10,240 × 10,240 px) of the expected performance are presented in Table A1.

These results were obtained on a Scan 3XS EGX server, with an AMD 32-core EPYC 7452, 256 GB RAM, and 4 × 48 GB NVIDIA Tesla A40 GPUs. Note that this server setup is different from that used in the original manuscript in order to permit parallel execution on four GPUs with a larger memory of 48 GB each. The original timing of 2 min (referred to in the Conclusion) for a single 2.5 × 2.5 km^2 was obtained on a server with an Intel i9-7960X CPU, 128 GB RAM, and 24 GB NVIDIA Titan RTX GPU.

The results indicated that with a single GPU, a patch of 50 × 50 km^2 can be computed in approximately 3 h 42 min. For an ROI of 100 × 100 km^2 and with adequate parallelisation, the same processing time was maintained. To further speed up the processing for time-sensitive applications, the patch size can be maintained as the original 2.5 × 2.5 km^2 and independently processed on more computing resources.

Table A1. The time and memory usage of the proposed technique. For 4 GPUs independently processing segmented patches, a tile as large as 100 km by 100 km can be processed in under 4 h.

Resources	1 GPU	1 GPU	1 GPU	1 GPU	2 GPUs	4 GPUs
ROI Size (km^2)	2.5 × 2.5	100 × 100	50 × 50	100 × 100	100 × 100	100 × 100
Patch Size (km^2)	2.5 × 2.5	2.5 × 2.5	50 × 50	50 × 50	50 × 50	50 × 50
# of Patches	1	1600	1	4	4	4
GPU RAM (MB)	1921	1921	38,655	38,655	38,655	38,655
Epoch Time (HH:MM:SS)	00:00:24	00:00:24	00:55:33	00:55:33	00:55:33	00:55:33
Total Time (HH:MM:SS)	00:01:35	42:14:02 *	03:42:09 *	14:48:41 *	07:24:21 *	03:42:09

* Quantities with asterisk (*) are extrapolated from a single epoch time for 4 epochs and for the entire ROI size. Single GPU mode processes each patch one after another.

References

1. Ali, A.M.; Savin, I.; Poddubskiy, A.; Abouelghar, M.; Saleh, N.; Abutaleb, K.; El-Shirbeny, M.; Dokukin, P. Integrated method for rice cultivation monitoring using Sentinel-2 data and Leaf Area Index. *Egypt. J. Remote Sens. Space Sci.* **2021**, *24*, 431–441. [CrossRef]
2. Sitokonstantinou, V.; Papoutsis, I.; Kontoes, C.; Lafarga Arnal, A.; Armesto Andrés, A.P.; Garraza Zurbano, J.A. Scalable Parcel-Based Crop Identification Scheme Using Sentinel-2 Data Time-Series for the Monitoring of the Common Agricultural Policy. *Remote Sens.* **2018**, *10*, 911. [CrossRef]
3. Gómez, D.; Salvador, P.; Sanz, J.; Casanova, J.L. Potato Yield Prediction Using Machine Learning Techniques and Sentinel 2 Data. *Remote Sens.* **2019**, *11*, 1745. [CrossRef]
4. Drusch, M.; Del Bello, U.; Carlier, S.; Colin, O.; Fernandez, V.; Gascon, F.; Hoersch, B.; Isola, C.; Laberinti, P.; Martimort, P.; et al. Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services. *Remote Sens. Environ.* **2012**, *120*, 25–36.
5. Roerink, G.J.; Menenti, M.; Verhoef, W. Reconstructing cloudfree NDVI composites using Fourier analysis of time series. *Int. J. Remote Sens.* **2000**, *21*, 1911–1917. [CrossRef]
6. Wang, Z.; Jin, J.; Liang, J.; Yan, K.; Peng, Q. A new cloud removal algorithm for multi-spectral images. In Proceedings of the MIPPR 2005: SAR Multispectral Image Process, Wuhan, China, 31 October–2 November 2005; Volume 6043, p. 60430W. [CrossRef]
7. Chen, J.; Zhu, X.; Vogelmann, J.E.; Gao, F.; Jin, S. A simple and effective method for filling gaps in Landsat ETM+ SLC-off images. *Remote Sens. Environ.* **2011**, *115*, 1053–1064. [CrossRef]
8. Mitchell, O.R.; Delp, E.J.; Chen, P.L. Filtering To Remove Cloud Cover in Satellite Imagery. *IEEE Trans. Geosci. Electron.* **1977**, *GE-15*, 137–141. [CrossRef]
9. Gabarda, S.; Cristóbal, G. Cloud covering denoising through image fusion. *Image Vis. Comput.* **2007**, *25*, 523–530. [CrossRef]
10. Holben, B.N. Characteristics of maximum-value composite images from temporal AVHRR data. *Int. J. Remote Sens.* **1986**, *7*, 1417–1434. [CrossRef]
11. Cihlar, J.; Howarth, J. Detection and Removal of Cloud Contamination from AVHRR Images. *IEEE Trans. Geosci. Remote Sens.* **1994**, *32*, 583–589. [CrossRef]
12. Wang, B.; Ono, A.; Muramatsu, K.; Fujiwarattt, N. Automated detection and removal of clouds and their shadows from landsat TM images. *IEICE Trans. Inf. Syst.* **1999**, *E82-D*, 453–460.
13. Li, M.; Liew, S.C.; Kwoh, L.K. Generating "cloud free" and "cloud-shadow free" mosaic for spot panchromatic images. *Int. Geosci. Remote Sens. Symp. (IGARSS)* **2002**, *4*, 2480–2482. [CrossRef]
14. Helmer, E.H.; Ruefenacht, B. Cloud-free satellite image mosaics with regression trees and histogram matching. *Photogramm. Eng. Remote Sens.* **2005**, *71*, 1079–1089. [CrossRef]
15. Champion, N. Automatic Cloud Detection From Multi-Temporal Satellite Images: Towards the Use of Pléiades Time Series. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2012**, *XXXIX-B3*, 559–564. [CrossRef]
16. Melgani, F. Contextual reconstruction of cloud-contaminated multitemporal multispectral images. *IEEE Trans. Geosci. Remote Sens.* **2006**, *44*, 442–455. [CrossRef]
17. Lorenzi, L.; Melgani, F.; Mercier, G. Missing-area reconstruction in multispectral images under a compressive sensing perspective. *IEEE Trans. Geosci. Remote Sens.* **2013**, *51*, 3998–4008. [CrossRef]
18. Lin, C.H.; Tsai, P.H.; Lai, K.H.; Chen, J.Y. Cloud removal from multitemporal satellite images using information cloning. *IEEE Trans. Geosci. Remote Sens.* **2013**, *51*, 232–241. [CrossRef]
19. Li, X.; Shen, H.; Member, S.; Zhang, L.; Member, S. Contaminated by Thick Clouds and Shadows Using Multitemporal Dictionary Learning. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 7086–7098.

20. Enomoto, K.; Sakurada, K.; Wang, W.; Fukui, H.; Matsuoka, M.; Nakamura, R.; Kawaguchi, N. Filmy Cloud Removal on Satellite Imagery with Multispectral Conditional Generative Adversarial Nets. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 1533–1541. [CrossRef]
21. Singh, P.; Komodakis, N. Cloud-GAN: Cloud removal for sentinel-2 imagery using a cyclic consistent generative adversarial networks. In Proceedings of the IGARSS 2018—2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018; pp. 1772–1775. [CrossRef]
22. Grohnfeldt, C.; Schmitt, M.; Zhu, X. A conditional generative adversarial network to fuse SAR and multispectral optical data for cloud removal from Sentinel-2 images. In Proceedings of the IGARSS 2018—2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018; pp. 1726–1729. [CrossRef]
23. Bermudez, J.D.; Happ, P.N.; Oliveira, D.A.; Feitosa, R.Q. SAR to Optical Image Synthesis for Cloud Removal with Generative Adversarial Networks. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2018**, *4*, 5–11. [CrossRef]
24. Rafique, M.U.; Blanton, H.; Jacobs, N. Weakly supervised fusion of multiple overhead images. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Long Beach, CA, USA, 16–17 June 2019; pp. 1479–1486. [CrossRef]
25. Bermudez, J.D.; Happ, P.N.; Feitosa, R.Q.; Oliveira, D.A. Synthesis of Multispectral Optical Images from SAR/Optical Multitemporal Data Using Conditional Generative Adversarial Networks. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 1220–1224. [CrossRef]
26. Gao, J.; Yuan, Q.; Li, J.; Zhang, H.; Su, X. Cloud removal with fusion of high resolution optical and SAR images using generative adversarial networks. *Remote Sens.* **2020**, *12*, 191. [CrossRef]
27. Meraner, A.; Ebel, P.; Zhu, X.X.; Schmitt, M. Cloud removal in Sentinel-2 imagery using a deep residual neural network and SAR-optical data fusion. *ISPRS J. Photogramm. Remote Sens.* **2020**, *166*, 333–346. [CrossRef] [PubMed]
28. Sarukkai, V.; Jain, A.; Uzkent, B.; Ermon, S. Cloud removal in satellite images using spatiotemporal generative networks. In Proceedings of the 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), Snowmass, CO, USA, 1–5 March 2020; pp. 1785–1794. [CrossRef]
29. Zhang, Q.; Yuan, Q.; Li, Z.; Sun, F.; Zhang, L. Combined deep prior with low-rank tensor SVD for thick cloud removal in multitemporal images. *ISPRS J. Photogramm. Remote Sens.* **2021**, *177*, 161–173. [CrossRef]
30. Ebel, P.; Meraner, A.; Schmitt, M.; Zhu, X.X. Multisensor Data Fusion for Cloud Removal in Global and All-Season Sentinel-2 Imagery. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 5866–5878. [CrossRef]
31. Maalouf, A.; Carré, P.; Augereau, B.; Fernandez-Maloigne, C. A bandelet-based inpainting technique for clouds removal from remotely sensed images. *IEEE Trans. Geosci. Remote Sens.* **2009**, *47*, 2363–2371. [CrossRef]
32. Shen, H.; Li, H.; Qian, Y.; Zhang, L.; Yuan, Q. An effective thin cloud removal procedure for visible remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2014**, *96*, 224–235. [CrossRef]
33. Sandhan, T.; Choi, J.Y. Simultaneous Detection and Removal of High Altitude Clouds from an Image. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 4789–4798. [CrossRef]
34. Salberg, A.B. Land Cover Classification of Cloud-Contaminated. *IEEE Trans. Geosci. Remote Sens.* **2011**, *49*, 377–387. [CrossRef]
35. Poggio, L.; Gimona, A.; Brown, I. Spatio-temporal MODIS EVI gap filling under cloud cover: An example in Scotland. *ISPRS J. Photogramm. Remote Sens.* **2012**, *72*, 56–72. [CrossRef]
36. Zeng, C.; Shen, H.; Zhang, L. Recovering missing pixels for Landsat ETM+ SLC-off imagery using multi-temporal regression analysis and a regularization method. *Remote Sens. Environ.* **2013**, *131*, 182–194. [CrossRef]
37. Zhu, X.; Gao, F.; Liu, D.; Chen, J. A modified neighborhood similar pixel interpolator approach for removing thick clouds in landsat images. *IEEE Geosci. Remote Sens. Lett.* **2012**, *9*, 521–525. [CrossRef]
38. Tseng, D.C.; Chien, C.L. A cloud removal approach for aerial image visualization. *Int. J. Innov. Comput. Inf. Control* **2013**, *9*, 2421–2440.
39. Wang, J.; Olsen, P.A.; Conn, A.R.; Lozano, A.C. Removing Clouds and Recovering Ground Observations in Satellite Image Sequences via Temporally Contiguous Robust Matrix Completion. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2754–2763. [CrossRef]
40. Huang, B.; Li, Y.; Han, X.; Cui, Y.; Li, W.; Li, R. Cloud removal from optical satellite imagery with SAR imagery using sparse representation. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 1046–1050. [CrossRef]
41. Chen, Y.; Tang, L.; Yang, X.; Fan, R.; Bilal, M.; Li, Q. Thick Clouds Removal from Multitemporal ZY-3 Satellite Images Using Deep Learning. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 143–153. [CrossRef]
42. Li, J.; Wu, Z.; Hu, Z.; Li, Z.; Wang, Y.; Molinier, M. Deep learning based thin cloud removal fusing vegetation red edge and short wave infrared spectral information for sentinel-2A imagery. *Remote Sens.* **2021**, *13*, 157. [CrossRef]
43. Ilyas, A.; Santurkar, S.; Tsipras, D.; Engstrom, L.; Tran, B.; Madry, A. Adversarial Examples Are Not Bugs, They Are Features. In Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 8–14 December 2019.
44. Ulyanov, D.; Vedaldi, A.; Lempitsky, V. Deep Image Prior. *Int. J. Comput. Vis.* **2020**, *128*, 1867–1888. [CrossRef]
45. Jayaram, V.; Thickstun, J. Source separation with deep generative priors. In Proceedings of the 37th International Conference on Machine Learning, Virtual Event, 13–18 July 2020.

46. Sidorov, O.; Hardeberg, J.Y. Deep hyperspectral prior: Single-image denoising, inpainting, super-resolution. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Korea, 27–28 October 2019; pp. 3844–3851. [CrossRef]
47. Zou, Z.; Lei, S.; Shi, T.; Shi, Z.; Ye, J. Deep Adversarial Decomposition: A Unified Framework for Separating Superimposed Images. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 12803–12813. [CrossRef]
48. Ebel, P.; Schmitt, M.; Zhu, X.X. Internal Learning for Sequence-to-Sequence Cloud Removal via Synthetic Aperture Radar Prior Information. In Proceedings of the 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, Brussels, Belgium, 11–16 July 2021; pp. 2691–2694. [CrossRef]
49. Zhang, Y.; Zhao, C.; Wu, Y.; Luo, J. Remote sensing image cloud removal by deep image prior with a multitemporal constraint. *Opt. Contin.* **2022**, *1*, 215–226. [CrossRef]
50. Mahajan, S.; Fataniya, B. Cloud detection methodologies: Variants and development—A review. *Complex Intell. Syst.* **2020**, *6*, 251–261. [CrossRef]
51. Upadhyay, P.; Czerkawski, M.; Davison, C.; Cardona, J.; Macdonald, M.; Andonovic, I.; Michie, C.; Atkinson, R.; Papadopoulou, N.; Nikas, K.; et al. A Flexible Multi-Temporal and Multi-Modal Framework for Sentinel-1 and Sentinel-2 Analysis Ready Data. *Remote Sens.* **2022**, *14*, 1120. [CrossRef]
52. Wang, Z.; Bovik, A.; Sheikh, H.; Simoncelli, E. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef]