This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2023.3333391

U-TILISE: A Sequence-to-sequence Model for Cloud Removal in Optical Satellite Time Series

Corinne Stucker, Vivien Sainte Fare Garnot, and Konrad Schindler, *Senior Member, IEEE*

***Abstract*—Satellite image time series in the optical and infrared spectrum suffer from frequent data gaps due to cloud cover, cloud shadows, and temporary sensor outages. It has been a long-standing problem of remote sensing research how to best reconstruct the missing pixel values and obtain complete, cloud-free image sequences. We approach that problem from the perspective of representation learning and develop U-TILISE, an efficient neural model that is able to implicitly capture spatio-temporal patterns of the spectral intensities, and that can therefore be trained to map a cloud-masked input sequence to a cloud-free output sequence. The model consists of a convolutional *spatial encoder* that maps each individual frame of the input sequence to a latent encoding; an attention-based *temporal encoder* that captures dependencies between those per-frame encodings and lets them exchange information along the time dimension; and a convolutional *spatial decoder* that decodes the latent embeddings back into multi-spectral images. We experimentally evaluate the proposed model on EarthNet2021, a dataset of Sentinel-2 time series acquired all over Europe, and demonstrate its superior ability to reconstruct the missing pixels. Compared to a standard interpolation baseline, it increases the PSNR by 1.8 dB at previously seen locations and by 1.3 dB at unseen locations.**

***Index Terms*—Image time series reconstruction, Self-attention, Cloud removal, Sentinel-2, Sequence-to-sequence model**

I. Introduction

MODERN satellite images have made it possible to continuously and systematically monitor the Earth's surface. Remotely sensed imagery, and products derived from it by image classification [1]–[3], segmentation [4], [5], or regression [6]–[8], have become an important data source for applications ranging from environmental monitoring [9]–[11] to agricultural management [12]–[14]. Moreover, multi-temporal satellite image sequences provide unprecedented opportunities to explore the temporal dynamics of natural processes, such as the evolution of land cover phenology [15]. Among Earth observation missions that systematically revisit the same locations and capture such sequences, most acquire images in the optical and near-infrared part of the electromagnetic spectrum. On the one hand, optical images are particularly suitable for visual interpretation by humans. On the other hand, spectral signatures in that part of the spectrum contain information that makes it possible to distinguish between different land cover types and to quantify the health and vitality of vegetation. Notably, several indicators for vegetation density and productivity are based on (heuristic, non-linear) combinations of spectral intensities between 400 nm and 1500 nm. Such indices, including the widely used normalized difference vegetation index (NDVI) [16], are a mainstay of current remote sensing practice [17], [18].

Unfortunately, the effective availability of optical satellite images is considerably lower than the nominal revisit frequency of the satellites. Data gaps routinely occur, either due to occlusions or because no image is captured during an overpass for technical reasons, such as sensor maintenance or conflicting imaging requests. The primary cause for data gaps is the weather, i.e., clouds, haze, and cloud shadows that partially or fully obscure the observed scene. A study over 12 years of optical data acquired by the MODIS sensor concluded that, on average, clouds occlude 67% of the Earth's surface and 55% of the land surface at any point in time [19]. The large proportion of data gaps, which moreover are irregularly distributed, calls for measures to ensure the usability of monitoring systems in the presence of frequent clouds.

A first, rudimentary approach adopted by several satellite processing pipelines is to discard images affected by clouds before further analysis. Processing solely cloud-free image observations may make data handling and visual inspection more convenient, but it also discards a lot of data that may still be usable, since often only a moderate fraction of an image or time series is affected by clouds. Moreover, [20] has shown that learning-based image classifiers trained on curated, cloud-free training datasets do not perform all that well when applied to images with even a small amount of clouds, let alone to images with moderate or severe cloud cover. More recent pipelines operate on all available input images, regardless of their degree of cloud cover, and learn to ignore uninformative pixels at an algorithmic level, for instance, via data-driven attention mechanisms [1]. An alternative strategy is to *remove* clouds and cloud shadows in advance, such that the subsequent processing is no longer affected by them, rather than (explicitly or implicitly) ignore them during image analysis. An advantage of such a two-step approach is that multiple image analysis pipelines tailored to different tasks can all use the same cloud-free images. These pipelines can then be more efficient and often also more robust, because they need not each devote a significant share of their model capacity to the detection and handling of data gaps [21].

In the last few years, encoder-decoder style neural networks have become the prevalent approach to recover missing data in optical satellite images. Several methods have tried to directly learn the mapping from a cloudy to a cloudless RGB image [22]–[24], others even try to translate Synthetic Aperture Radar (SAR) images to multi-spectral optical images [25]–[29]. A recent trend has been to combine the two inputs and perform SAR-optical data fusion [30]–[32]. SAR systems

Corinne Stucker and Konrad Schindler are with the Chair of Photogrammetry and Remote Sensing, ETH Zurich, 8093 Zurich, Switzerland (e-mail: stuckerc@ethz.ch; schindler@ethz.ch).

Vivien Sainte Fare Garnot is with the Institute for Computational Science, University of Zurich, 8057 Zurich, Switzerland (e-mail: vsaint@ics.uzh.ch).

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2023.3333391

employ active illumination with much longer wavelengths, which are unaffected by clouds and shadows and may provide complementary information to guide the reconstruction of occluded content. However, one must bridge a considerable domain gap to infer optical reflectance values from SAR amplitudes. Both are likely to change at land cover boundaries, so SAR may help to restore image gradients, but it can hardly be expected to offer much information about the actual spectral values, colors, and texture details. In contrast, sequences of optical images depicting the same location[1] exhibit strong correlations along the temporal dimension. These spatio-temporal patterns provide evidence about the spatial structure *and* the spectral properties. Importantly, this holds not only for static or gradually changing land cover but also includes more complex temporal dynamics, like seasonal variations. The idea to leverage multi-temporal images for cloud removal is rather obvious; but most existing approaches [31], [33]–[36] collapse the multi-temporal, cloudy input into a single gap-free image, often additionally relying on spatially and temporally co-registered SAR observations to guide the reconstruction [31].

Here, we focus on full *sequence-to-sequence* translation: our goal is to convert the cloudy input time series into a gap-free product with the same time steps, but containing a clean, cloud-free image at every frame (no matter whether the original frame was cloud-free, cloudy, or entirely missing). To that end, we introduce U-TILISE,[2] a neural image sequence model that captures spatio-temporal relationships between the spectral intensities in an image time series, and is therefore able to impute missing pixels. U-TILISE operates in three dimensions, with 2D convolutions to encode multi-scale local relationships in space and 1D (self-)attention to encode non-local relations in time. By design, its output has the same spatial and temporal extent as the input, such that it jointly reconstructs complete, gap-free time series. The model supports additional, auxiliary input channels and is therefore, in principle, able to use SAR amplitudes, too. But empirically this brings only a tiny improvement (see appendix).

We experimentally test U-TILISE on the EarthNet2021 dataset [37], which contains thousands of 30-frame, 4-channel (R, G, B, NIR) time series of Sentinel-2 data. Compared to standard interpolation between the temporally nearest unoccluded observations, our model improves the peak signal-to-noise ratio (PSNR) of the reconstructed spectral values by 1.8 dB at previously observed locations, and by 1.3 dB at unseen locations.

The remainder of this article is organized as follows: first, we provide an overview of cloud removal methods in Section II, with a focus on learned (mono-temporal as well as sequence-based) approaches. In Section III, we introduce the U-TILISE model, set out its components (Sections III-C and III-D), and describe the associated training and inference procedures (Section III-E). Next, we explain the data (Section IV) and the experimental setup (Section V) used in our evaluation. Section VI reports and discusses experimental results, followed by a conclusion (Section VII). Complementary experiments with additional SAR observations on top of optical time series are given in the appendix.

[1]Throughout this paper, the shorthand "time series" denotes a temporally ordered, co-registered sequence of images showing the same location.

[2]U-Net Temporal Imputation Lightweight Image Sequence Encoder

II. Related Work

Reconstructing missing pixels in remotely sensed imagery has been a long-standing research problem. Early efforts towards thin cloud and haze removal build on physical relations [38]–[40] or signal processing considerations [41], [42] to describe the process of light transmission and interaction with clouds. Methods designed to recover image content occluded by thick clouds have been based on tensor factorization [43]–[45], on mosaicking of multi-temporal images [46]–[49] or they adopt statistical image processing methods originally developed for single-image inpainting [50]–[52]. In the following, we concentrate on data-driven, learning-based methods for cloud removal. For completeness, we note that video inpainting methods like [53] share conceptual similarities with satellite time series imputation, but they are beyond the scope of the present literature review.

A. Mono-temporal cloud removal

A natural formulation of cloud removal is as an image-to-image translation task, where the mapping from the cloudy input to the cloud-free output is learned in a data-driven manner. For instance, [22] employ a conditional generative adversarial network (cGAN) [54] to map from a cloudy to a cloudless RGB satellite image. The mapping is conditioned on the NIR channel of the input, arguing that near-infrared wavelengths partially penetrate clouds and may thus capture information about the observed scene that is unavailable in the visible spectrum. In [23], the NIR channel is replaced with conditioning on a SAR image, as clouds are completely transparent at radar wavelengths. The works of [22], [23] both do not go beyond a proof-of-concept; the underlying neural networks are trained and evaluated exclusively on synthetically generated images with clouds simulated by Perlin noise [55]. It has since been shown that such simulations generalize poorly to real cloudy images [56]. To side-step the need for training examples where cloudy and cloud-free images are in exact, pixel-wise correspondence, [24], [56] rely on a cycle-consistency loss [57]. In this way, the networks can be trained directly on images with real data gaps, eliminating the potential domain gap between training and test data. While the method in [24] is limited to thin clouds, [56] do not impose any restrictions on the maximum permissible cloud coverage or density. Furthermore, [56] combine explicit modeling of cloud densities with a residual learning strategy to better preserve the pixel values in cloud-free image regions. The methods mentioned so far have in common that they are limited to images with three optical bands. To address that limitation, several SAR-to-optical image translation approaches [25]–[29] learn the mapping from a SAR image to the full stack of multi-spectral bands, often also using cGANs.

Recent advances in learned cloud removal tend to rely on image fusion, i.e., they synergistically use the cloudy optical and a cloud-free SAR image to impute the missing pixels in

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2023.3333391

the former. In [30], a Sentinel-2 image and a temporally close SAR image of the same scene are stacked together along the channel dimension and fed into a neural network that regresses a residual reflectance value at every pixel. Those per-pixel corrections are then added to the input to remove the missing data. [58] combine SAR-to-optical translation and SAR-optical data fusion in a cascaded fashion. First, a GAN is trained to map the SAR input to an optical image. That synthetic image is stacked with the original SAR data and the cloudy optical input and fed into a second GAN, trained to map the multi-modal input to a cloud-free optical image.

The recent [32] found that stacking optical and SAR observations into a multi-modal image and processing them together does not optimally exploit the two inputs, as feature extraction from SAR is aggravated by speckle noise. Instead, the authors propose separate embedding branches per modality, together with an attention-based mechanism that gradually and selectively fuses features from the two branches.

SAR-optical data fusion approaches have demonstrated that complementary information in the form of SAR observations can help to compensate for missing data in optical images. Still, fusing optical and SAR data remains challenging due to the large domain gap between the two modalities. One must also keep in mind that SAR can hardly contribute to restoring actual spectral reflectance information, like different hues or fine-grained textures. Its role is to add spatial context, such as land-cover boundaries, which appear as gradients also in the SAR amplitude. An alternative approach is to inject contextual information from other optical sensors, as, for instance, in [59]. Clearly, this will greatly reduce the domain gap, but on the other hand, there is no guarantee that a temporal close and largely cloud-free image can be found.

B. Sequence-based cloud removal

Sequence-to-point methods [31], [33]–[36], [60], [61] consume a multi-spectral time series with data gaps and output a single, gap-free image. In most cases, that output has not got a well-defined time stamp but rather is seen as representative of the entire time period between the start and end dates of the time series. Even if the output is associated with a specific time, e.g., the middle frame, the method would have to be run iteratively to reconstruct an entire time series. Typically, the input time series only have three to five images. Optionally, the reconstruction can additionally be guided by a single SAR image [61] or by a SAR time series (approximately) aligned frame-to-frame with the input [31]. Some of these sequence-to-point approaches impose tight restrictions on the maximum cloud coverage per image, e.g., [33] require at most 10–30% cloud cover, and [60] require the first and last of three frames to be cloud-free (0% cover). Furthermore, these methods often assume short temporal intervals with minimal land cover changes over time, since they accumulate spectral information along the temporal dimension to create the output image. That assumption rules out systematic land cover dynamics, in particular seasonal cycles of vegetation and agriculture.

To our knowledge, [62]–[64] are the only published sequence-to-sequence models, i.e., they output a cloud-free, multi-spectral image for every frame of the input time series. In [63], the sequence model is parameterized as a recurrent neural network with a two-layer GRU [65] architecture and used to learn the mapping from a SAR time series to a time series of Landsat images, only for pixels belonging to a specific land cover class, namely rice fields. This yields reconstructions of limited quality (PSNR $< 28\,\mathrm{dB}$), possibly due to the well-known difficulties of learning multi-layer recurrent models [66]. A rather different approach is taken by [62], who adopt a recent video inpainting technique [67] that extends the deep image prior (DIP) [68] to videos. The initial input is a SAR time series (as opposed to random noise in the original DIP) such that the network effectively performs a mapping from SAR to optical time series. We point out that while the DIP employs a neural network parameterization, it is not a learned model. The convolutional network structure, which favors lower amplitudes for high-frequency signals, serves as a hard-wired low-level prior of optical image statistics. It does not store any a priori information extracted from training data. Instead, the network weights are optimized individually for every input sequence at inference time. The method most related to our proposed U-TILISE model is [64], an adversarial approach that internally splits the computation into a first, coarse round of imputation and a subsequent refinement network. The coarse imputation model is a 3D spatio-temporal encoder-decoder architecture with separate backbones for the optical and SAR inputs, followed by a transformer-style attention mechanism in the bottleneck to fuse the latent embeddings of the two encoder branches. The model does not implicitly learn to ignore cloudy observations; instead, the contribution of cloudy input pixels is suppressed by explicitly modulating the learned attention masks according to the given cloud masks before applying them to the latent embeddings coming out of the optical encoder branch.

III. Method

To impute missing image content, we design U-TILISE, a learnable image sequence model in the form of a neural network. Once trained, the weights of that network encode a prior over spatio-temporal patterns of multi-spectral reflectance. When a time series with data gaps is fed into the network, the prior fills in the missing values to obtain a complete, gap-free time series. In contrast to existing cloud removal pipelines, our approach does not rely on auxiliary SAR observations to guide the imputation. Instead, we exploit spatial and temporal patterns within the multi-spectral time series itself to reconstruct the spatio-temporal evolution of the depicted land cover. Furthermore, our model jointly reconstructs all images in a given time series, as opposed to pipelines that reconstruct a single frame, considered representative of the entire sequence.

A. Problem formulation

Let $\boldsymbol{X} \in \mathbb{R}^{T \times C \times H \times W}$ denote a multi-spectral time series, represented as a four-dimensional tensor with T the temporal length, C the number of spectral bands, and $H \times W$ the spatial extent. Our goal is to regress a reflectance for every spatio-temporal location, so as to obtain a complete, gap-free multi-spectral time series $\hat{\boldsymbol{Y}} \in \mathbb{R}^{T \times C \times H \times W}$. Our model assumes

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2023.3333391

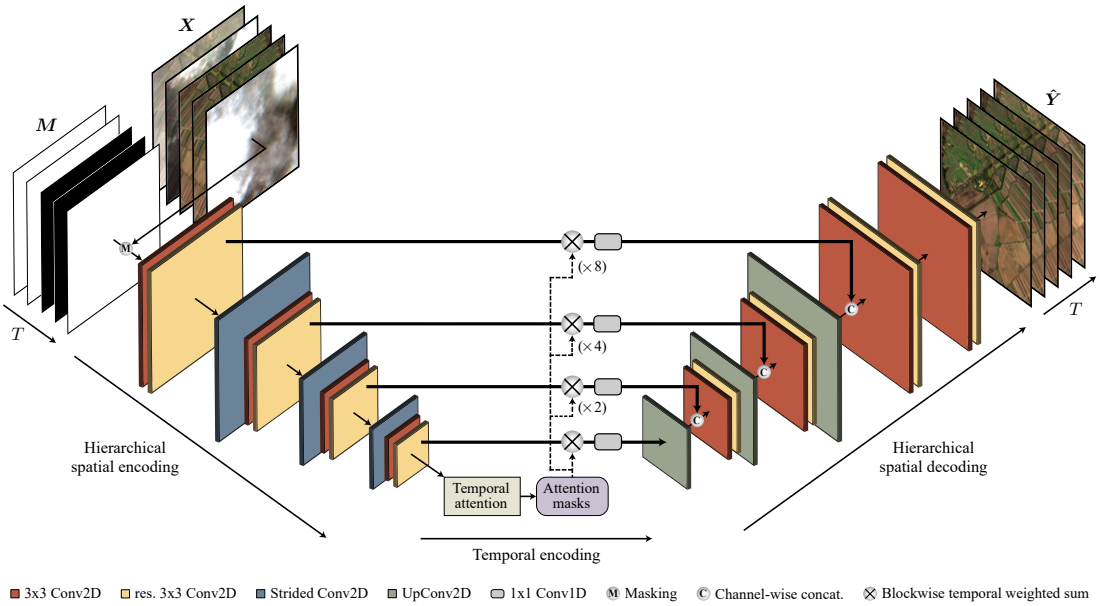

Fig. 1. Overview of the proposed model: U-TILISE is a neural sequence-to-sequence model that takes as input a multi-spectral satellite image time series in which missing reflectance values are masked and outputs a complete and gap-free time series with the same dimensions. U-TILISE employs a convolutional U-Net architecture [69] over the spatial and spectral dimensions and a transformer-style self-attention mechanism along the temporal dimension. The attention masks operate on the (spatial) bottleneck between the encoder and decoder parts of the U-Net as well as on the skip connections (after suitable upsampling).

that the spatio-temporal locations to be imputed are marked by a binary mask $\boldsymbol{M} \in \mathbb{R}^{T\times 1\times H\times W}$, where the value 1 denotes pixels with a valid observation and 0 denotes missing data values. Note that we do not impose any assumptions or requirements on the mask $\boldsymbol{M}$: it may denote any type of data gaps, including clouds and cloud shadows, but also frames that are entirely missing, for instance, due to sensor maintenance.

B. Overview

Fig. 1 gives a graphical overview of our method. At its core is U-TILISE, a learned sequence-to-sequence model that captures spatio-temporal relationships between the spectral intensities, and thereby is able to reconstruct realistic, complete, and gap-free multi-spectral time series. For efficiency, the mask $\boldsymbol{M} \in \mathbb{R}^{T\times 1\times H\times W}$ of missing pixels is not separately fed into the model but imprinted directly on the multi-spectral input $\boldsymbol{X} \in \mathbb{R}^{T\times C\times H\times W}$ by setting all masked pixels to the maximum intensity 1. U-TILISE consists of three components. First, a shared multi-scale convolutional encoder transforms every image of the masked sequence into a latent embedding. Next, an attention-based temporal encoder combines the per-frame embeddings across time to impute missing values in the latent sequence representation. Last, a shared convolutional decoder reconstructs every image from its latent embedding to obtain a gap-free time series with the spatial, spectral, and temporal dimensions of the input.

C. 3D spatio-temporal sequence-to-sequence model

U-TILISE builds on recent advances in learned time series processing. Its architecture is inspired by U-TAE [4], a model originally developed for crop mapping, which maps a time series to a (mono-temporal) panoptic segmentation. In a nutshell, U-TAE combines convolutions for multi-scale spatio-spectral encoding with a lightweight non-local temporal attention mechanism [2]. Intuitively, the latter learns to focus on the most salient observations within a given time series. By design, U-TAE collapses the input along the temporal dimension to produce a mono-temporal output. We take inspiration from the design of modern transformer models [70], [71] and extend the architecture to a full 3D spatio-temporal sequence-to-sequence model that preserves the temporal dimension.

U-TILISE consists of *(i)* symmetric multi-scale spatio-spectral encoding and decoding modules in the style of U-Net [69] and *(ii)* a lightweight temporal encoding module based on multi-head self-attention [72], see Fig. 1. We now describe each of these components in more detail.

1) Spatial encoder: The spatial encoder gradually transforms the masked time series of size $(T \times C \times H \times W)$ into a multi-scale latent embedding via a sequence of convolutional blocks. Each block comprises a 3×3 convolutional layer with stride 1 and d filter channels, followed by a rectified linear unit (ReLU) as non-linear activation function and a residual 3×3 convolution with stride 1, d' filter channels, and ReLU activation. Between the convolutional blocks, a strided convolution equipped with ReLU activation decreases the spatial resolution of the intermediate embeddings by a factor of 2. After encoding all images in the time series (individually and in parallel), we temporally stack their latent representations to produce a multi-temporal sequence embedding with dimensions $(T \times D \times H/8 \times W/8)$, with D the channel depth of that embedding.

2) Temporal encoder: The temporal encoder operates individually on the spatial locations (low-resolution "pixels") of the latent embedding. For each such pixel, it captures the pairwise dependencies between the values in all pairs of different

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2023.3333391

frames and uses them to fill in missing information. The temporal encoder is based on the Lightweight Temporal Attention Encoder (L-TAE) of [2], which, in turn, is a simplified version of the multi-head self-attention mechanism of the transformer architecture [72]. Unlike [2], we employ data-driven queries to preserve the temporal dimension of the input. Moreover, we use residual skip connections as in the original transformer model [72]. We retain the computational simplifications of [2] and use a channel grouping, where the G attention heads process mutually exclusive subsets of D/G channels of the embedding. The learned attention scores are directly applied to the embedding vectors that come from the encoder (without first modulating them with a fully-connected layer). Following recent findings about neural sequence-to-sequence models, we prefer pre-normalization with the *groupnorm* scheme [73]. Furthermore, we employ GELU activations [74] rather than the classical ReLU activations in the multi-layer perceptron (MLP) of the attention block.

3) Spatial decoder: After the latent representation has been passed through the attention module, the spatial decoder progressively restores multi-spectral images from the individual per-image embeddings. These images have the same spectral and spatial resolution as the input to the network but no more missing values. The structure of the spatial decoding blocks is the same as for the spatial encoding blocks, except that fractionally strided, transposed convolutions with stride $\frac{1}{2}$ replace the strided convolutions. Once the native spatial resolution of the input has been reached, a final convolutional block maps the latent embedding to the spectral space. The final layer uses sigmoid activations instead of ReLU, so as to regress reflectances in the range $[0, 1]$. Finally, the reconstructed frames are stacked along the temporal dimension to recover the complete, gap-filled time series.

4) Skip connections: Skip connections from encoder to decoder levels of equal spatial resolution are a key component of the U-Net [69] architecture to propagate high-frequency details and localization information that is lost during spatial downsampling operations. We adopt the same strategy as in [4] and temporally weight the information transferred between corresponding layers of the spatial encoder and decoder. The attention masks learned by the temporal encoder serve as weights, which we spatially upsample to the adequate spatial resolution using bilinear interpolation. The temporally weighted output of the encoding layers is processed with a shared 1×1 convolutional layer followed by ReLU activation before channel-wise concatenation with the output of the corresponding decoding layers for further processing.

D. Sinusoidal positional encoding

By itself, the self-attention mechanism is agnostic to the sequence order. To provide positional information, we follow the standard procedure for transformers [72] and add a positional encoding (PE) to the input of the temporal encoder before applying self-attention:

$$\mathrm{PE}\,(t, k) = \sin\left(\mathrm{day}(t)/\tau^{\frac{2k}{D}} + \frac{\pi}{2} \bmod (k, 2)\right). \qquad (1)$$

$\mathrm{PE}\,(t, k)$ consists of fixed sinusoidal functions with predefined wavelengths and describes the position of the tth observation in the sequence, with D the channel depth of the embedding and k the coordinates of the positional encoding. We set $\tau = 1\,000$, as in [2]. Contrary to [72], we do not directly encode the ordinal position t in the sequence. Instead, we encode the observation date $\mathrm{day}(t)$, expressed as the number of days since the 1st of January of the respective calendar year. This strategy has proved beneficial for learned time series processing [2], [75], since it preserves information about seasonal patterns (e.g., lighting conditions or phenology of the vegetation) and accounts for irregular temporal sampling.

E. Training and inference

It is not possible to quantitatively assess the performance of time series imputation for real data gaps due to the lack of ground truth reflectances. Therefore, we train and evaluate U-TILISE with simulated data (generated by masking cloud-free frames with real cloud masks taken from other sequences, cf. Section IV-C) and examine its capability to generalize to sequences with actual data gaps due to real clouds. We train U-TILISE in a supervised manner by minimizing the pixel-wise absolute differences between the imputed time series $\hat{\boldsymbol{Y}}$ and the corresponding ground truth time series $\boldsymbol{Y}$:

$$\mathcal{L}_1 = \frac{1}{NCHW} \sum_{i=c=x=y=1}^{N,C,W,H} \frac{1}{T^i} \sum_{t=1}^{T^i} \mid \hat{Y}^i_{tcxy} - Y^i_{tcxy} \mid, \qquad (2)$$

where N denotes the number of training sequences and T^i the length of the ith sequence.

We train U-TILISE for a fixed temporal length of $T = 10$, i.e., the input is a time series that comprises at most 10 images. Shorter sequences are padded with no-data frames. During training, longer sequences are randomly cropped if $T^i > T$. At test time, we retain the full time series and process it in one shot if $T^i \leq T$, or in sliding window fashion if $T^i > T$.

IV. Data

We evaluate our method on EarthNet2021, a large, publicly available dataset (cf. Table III) of Sentinel-2 time series. Note, however, that our method is sensor-agnostic and can adapt to the properties of different multi-spectral imaging sensors, given suitable training data.

A. EarthNet2021 dataset

The EarthNet2021 [37] dataset was originally designed for satellite image forecasting, conditioned on future meteorological variables. It includes more than 32 000 Sentinel-2 time series collected over Central and Western Europe from November 2016 to May 2020. Each time series consists of 30 images with Level-1C top-of-atmosphere (TOA) reflectances. The images are acquired in a regular temporal interval of five days, where acquisition dates without an observation are encoded as images of NaN values. Every image is composed of the four spectral bands B2 (blue), B3 (green), B4 (red), and B8 (near-infrared) and covers a spatial extent of 128×128 pixels (2.56×2.56 km in scene space), resampled

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2023.3333391

to the resolution of 20 m. For every observation, the dataset further includes a pixel-wise cloud probability map[3] obtained via the `S2Cloudless` algorithm [76] and a binary cloud and cloud shadow mask based on heuristic rules similar to [30].

In our experiments, we reserve $\approx 20\%$ of the training sequences for validation, where training and validation tiles are mutually exclusive. For testing, we use the *iid* and *ood* test splits. Time series in the *iid* test split stem from the same Sentinel-2 tiles as the training data and the ones in the *ood* test split from previously unseen locations.

B. Preprocessing

We adopt the preprocessing protocol of prior works [37] and value-clip the optical images to the range $[0, 10\,000]$, followed by normalization to the unit range of $[0, 1]$.

To train a system for cloud removal that regresses time series rather than a single image, we found experimentally that pixel-wise supervision for every spatio-temporal location is crucial for learning seasonal changes and land surface dynamics over time. Since obtaining such ground truth for time series with real data gaps is impossible, we resort to cloud-free time series and introduce synthetically generated data gaps during training and evaluation, as described in Section IV-C. Starting from a time series with real data gaps, we first identify all images with partially occluded pixels or images that are occluded/missing entirely by applying a threshold to the cloud probability maps (if available) or the binary cloud masks. We choose the threshold in a conservative manner to minimize the number of undetected data gaps. We then remove all images with data gaps to produce cloud-free time series that exhibit a valid[4] observation for every spatio-temporal location. Second, we discard time series with less than five remaining images, as we deem such sequences too short for learning spatio-temporal patterns. Table III summarizes the number of time series, their temporal lengths, and the temporal resolution before and after filtering images with data gaps.

C. Simulation of data gaps

Realistically simulating cloud cover in satellite images is notoriously difficult. Synthetic images generated with existing physics-inspired simulation methods, like the well-known Perlin noise model [55], do not match the radiometry of real data well enough: it has been shown that cloud removal methods trained with such synthetic images do not generalize well to images containing actual clouds [56]. To create time series with artificial data gaps for training and evaluation, we thus refrain from rendering synthetic clouds. Instead, we adopt a strategy commonly employed for image inpainting [77]–[79] and completely mask out invalid pixels by setting them to the maximum intensity value 1. In this way, one only has to realistically simulate binary cloud masks, which is straightforward: all one needs to do is randomly sample real cloud masks from other acquisition times and/or locations within the same Sentinel-2 tile and apply them to a gap-free time series. With that strategy, we obtain time series with data gaps of realistic shapes and sizes and with known ground truth reflectances at all masked pixels in a controlled, fully automatic manner.

[3]Pixel-wise cloud probabilities are only available for the training data.

[4]A few cloudy images may remain in the cloud-filtered time series due to inaccuracies of the preceding cloud detector.

V. Experiments

A. Setup

Sequence-to-point methods typically use a temporally close cloud-free image of the same location to quantify the fidelity of the synthesized output. [31], one of the few existing sequence-to-sequence approaches, regresses a time series of multi-spectral Sentinel-2 images but restricts the evaluation to a single time step, namely the one with the lowest cloud cover. Such single-frame evaluation protocols are, in our view, problematic. On the one hand, they do not actually measure the quality of the regressed time series, as the temporal aspect is completely ignored. On the other hand, metrics computed only from the least cloudy image will likely be too optimistic, since the imputation task becomes more challenging with increasing occlusions.

We argue that the evaluation should take into account all images with missing pixels, irrespective of the degree of cloud cover. As ground truth reflectances are unavailable for time series with real data gaps, we quantitatively evaluate U-TILISE on cloud-free sequences with synthetically added data gaps. Additionally, we apply the learned model (without further fine-tuning) to time series featuring real data gaps to qualitatively assess performance in the true application setting.

We follow the procedure described in Sections IV-B and IV-C to generate time series with synthetic data gaps. Unless stated otherwise, we randomly trim the cloud-filtered time series to a maximum length of $T = 10$ images during training. At test time, we process the full-length sequences in sliding window fashion if their length exceeds T. To simulate data gaps, we randomly superimpose at most 50% of the images per time series with cloud (and cloud shadow) masks drawn randomly from the dataset, with a minimum of one masked image per sequence. When processing sequences with real data gaps, the actual cloud masks are used to mark missing pixels.

B. Implementation details

U-TILISE is implemented in PyTorch [80]. For training, we use an NVIDIA GeForce RTX 2080 Ti GPU for time series with four spectral bands and an NVIDIA TITAN RTX GPU for time series with 13 spectral bands. Source code and pretrained models are available at https://github.com/prs-eth/U-TILISE.

In our experiments, we use 64 filter channels for the spatial encoding and decoding convolutional layers, except at the lowest spatial resolution, where we use 128 filter channels. Accordingly, the temporal encoder has a latent feature dimension of 128. Temporal self-attention employs $G = 4$ heads, with a dimension of four for the data-driven keys and queries. To augment the training data, we randomly rotate all images in a sequence by $\alpha \in \{0°, 90°, 180°, 270°\}$ and randomly flip them along the x and y axes.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2023.3333391

We train with the Adam optimizer [81] with hyper-parameters $\{\beta_1$=0.9, β_2=0.999$\}$, a batch size of three, and no weight decay. During the first 250 epochs, the initial base learning rate of $2 \cdot 10^{-4}$ is halved every 50 epochs. Training is stopped once the $\mathcal{L}_1$ loss (cf. Section III-E) on the validation set has converged. In our experiments, this took about 1 000 epochs, or 20 days of training on a single GPU. The computational cost for applying the trained model is low: The forward pass for a time series consisting of (at most) 10 images takes $\approx$0.02 seconds. Longer sequences are processed in sliding window fashion, which on average takes 0.13 seconds for a 30-frame sequence.

C. Evaluation metrics

We adopt a suite of metrics commonly used to evaluate cloud removal and inpainting methods: mean absolute error (MAE), root mean square error (RMSE), spectral angle mapper (SAM) [82], peak signal-to-noise ratio (PSNR), and structural similarity index (SSIM) [83]:

$$\text{MAE} = \frac{1}{C\,|\Omega^i|} \sum_{c=1}^{C} \sum_{(t,x,y)\in\Omega^i} |\,\hat{Y}^i_{tcxy} - Y^i_{tcxy}\,|, \tag{3}$$

$$\text{RMSE} = \sqrt{\frac{1}{C\,|\Omega^i|} \sum_{c=1}^{C} \sum_{(t,x,y)\in\Omega^i} \left(\hat{Y}^i_{tcxy} - Y^i_{tcxy}\right)^2}, \tag{4}$$

$$\text{SAM} = \frac{1}{|\Omega^i|} \sum_{(t,x,y)\in\Omega^i} \arccos\left(\frac{\hat{\boldsymbol{Y}}^i_{txy} \cdot \boldsymbol{Y}^i_{txy}}{\|\hat{\boldsymbol{Y}}^i_{txy}\| \cdot \|\boldsymbol{Y}^i_{txy}\|}\right), \tag{5}$$

$$\text{PSNR} = 20 \cdot \log_{10}\left(\frac{1}{\text{RMSE}}\right), \tag{6}$$

$$\text{SSIM} = \frac{1}{|T^i_m|} \sum_{t\in T^i_m} \frac{\left(2\mu_{\hat{\boldsymbol{Y}}^i_t}\mu_{\boldsymbol{Y}^i_t} + \epsilon\right)\left(2\sigma_{\hat{\boldsymbol{Y}}^i_t \boldsymbol{Y}^i_t} + \epsilon'\right)}{\left(\mu^2_{\hat{\boldsymbol{Y}}^i_t} + \mu^2_{\boldsymbol{Y}^i_t} + \epsilon\right)\left(\sigma^2_{\hat{\boldsymbol{Y}}^i_t} + \sigma^2_{\boldsymbol{Y}^i_t} + \epsilon'\right)}, \tag{7}$$

where $\hat{\boldsymbol{Y}}^{\boldsymbol{i}}$ denotes the ith predicted time series with C spectral bands and T^i frames, $\boldsymbol{Y}^i$ the corresponding ground truth time series, and (x, y) and t the spatial and temporal coordinates. In the SSIM computation, $\mu_{\hat{\boldsymbol{Y}}^i_t}$ and $\sigma_{\hat{\boldsymbol{Y}}^i_t}$ express the mean reflectance and standard deviation of $\hat{\boldsymbol{Y}}^i$ at time step t, $\mu_{\boldsymbol{Y}^i_t}$ and $\sigma_{\boldsymbol{Y}^i_t}$ the corresponding mean and standard deviation w.r.t. $\boldsymbol{Y}^i$, and $\sigma_{\hat{\boldsymbol{Y}}^i_t \boldsymbol{Y}^i_t}$ their covariance. ϵ and ϵ' are small constants for numerical stability.

MAE, RMSE, and PSNR are popular metrics that quantify the pixel-wise reconstruction error. MAE and RMSE are expressed relative to the TOA reflectance range (ρ_{TOA}, recall that reflectances have been rescaled $\rho_{\text{TOA}} \in [0, 1]$) PSNR is in decibel (dB). SAM measures the spectral fidelity of the reconstructed pixels, defined as the average angle (in degrees) between predicted and ground truth spectral vectors. SSIM is a unit-less per-image metric that measures the overall structural similarity between a reconstructed image and the corresponding ground truth. We compute the pixel-based metrics over all imputed pixels (according to the input mask). Similarly, to compute the average SSIM, we only take images into account that contain masked pixels, denoted as T^i_m.

To generate the cloud-free reference time series for evaluation, we define a global threshold on the cloud probability maps (if available) or the binary cloud masks. That threshold may not be ideal for every single image; consequently, small clouds or haze may remain undetected. To alleviate the impact of these remaining data gaps (in the ground truth) on quantitative analysis, we compute the pixel-based metrics only for those imputed pixels (t, x, y) for which a cloud-free reference reflectance is available according to the original cloud masks, denoted as Ω^i for the ith time series.

D. Baseline methods

We compare U-TILISE against several natural baselines that operate independently on every spatial location over time. The simplest baseline (*last*) imputes missing pixels by copying the last valid observation before the current frame. The next baseline (*closest*) copies either the last or the next following observation, depending on which one is closer in time. The third baseline linearly interpolates between the last and next observations (according to the absolute time span in days), an approach frequently employed in operational practice [84].

VI. Results

A. Imputation quality of U-TILISE

We begin by evaluating our method on Sentinel-2 time series augmented with synthetic data gaps. Quantitative results are shown in Table I, visual examples are given in Fig. 2 and Fig. 3. U-TILISE generates coherent and gap-free optical time series that capture the natural evolution of the depicted land cover. It can handle occlusions of various shapes and sizes, recover images that suffer from severe occlusions, and reconstruct complete times series from input sequences that include multiple consecutive frames of missing data. Note how the model has implicitly learned to adapt to radiometric variations within the time series,[5] such that imputed pixels seamlessly blend into the surrounding image content, even if the frames that show the same regions unoccluded are radiometrically different. Besides naturally adapting reflectance values, U-TILISE is able to recover plausible color transitions and impute missing information in scenes with non-trivial temporal dynamics (Fig. 2, row 5). Furthermore, it preserves high reflectance values not associated with clouds (Fig. 3, row 3). Abrupt scene changes, such as harvested agricultural fields, are sometimes missed (Fig. 3, row 7), which is natural since their exact timing depends on weather conditions that the model has no access to.

Quantitatively, the predicted reflectances agree well with the true values across all optical bands. See Fig. 4. U-TILISE yields a MAE below 1% of the intensity range and a PSNR on the order of 38 dB. Recall, these errors are averaged only over imputed pixels and not inflated by the trivial reconstruction of observed values (cf. Table I, *iid* test split). The SSIM, averaged over all images with data gaps, amounts to 0.97. Notably, there is only a moderate performance penalty when applying the model to previously unseen locations (cf. Table I, *ood* test split). The MAE increases by 27% to roughly 0.01 and the SAM by 10% from 1.9 degrees to 2.1 degrees. Yet, with a

[5] Primarily caused by illumination changes due to atmospheric effects.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2023.3333391

TABLE I

QUANTITATIVE COMPARISON OF DIFFERENT IMPUTATION METHODS FOR TIME SERIES OF THE EARTHNET2021 DATASET, SEPARATELY EVALUATED FOR GEOGRAPHIC LOCATIONS SIMILAR TO THE TRAINING SET (*iid*) AND FOR PREVIOUSLY UNSEEN LOCATIONS (*ood*). COLUMNS 3–7 REPORT METRICS COMPUTED OVER ALL PIXELS (OR IMAGES, IN THE CASE OF SSIM) WITH MISSING INPUT DATA, WHEREAS COLUMNS 8–9 MEASURE THE QUALITY OF OUTPUT PIXELS (OR IMAGES) WITH VALID INPUT VALUES.

Test split	Method	↓ MAE [ρ_{TOA}]	↓ RMSE [ρ_{TOA}]	↓ SAM [°]	↑ PSNR [dB]	↑ SSIM [-]	↓ MAE [ρ_{TOA}]	↑ SSIM [-]
iid	Last	0.0148	0.0234	3.17	33.90	0.944	**0.0000**	**1.000**
	Closest	0.0128	0.0203	2.74	35.01	0.953	**0.0000**	**1.000**
	Linear interpolation	0.0110	0.0172	2.35	36.48	0.962	**0.0000**	**1.000**
	U-TILISE (ours)	**0.0086**	**0.0140**	**1.87**	**38.29**	**0.970**	0.0002	**1.000**
ood	Last	0.0174	0.0270	3.36	33.13	0.936	**0.0000**	**1.000**
	Closest	0.0147	0.0230	2.84	34.31	0.948	**0.0000**	**1.000**
	Linear interpolation	0.0129	0.0197	2.42	35.67	0.957	**0.0000**	**1.000**
	U-TILISE (ours)	**0.0110**	**0.0170**	**2.07**	**36.95**	**0.964**	0.0002	**1.000**

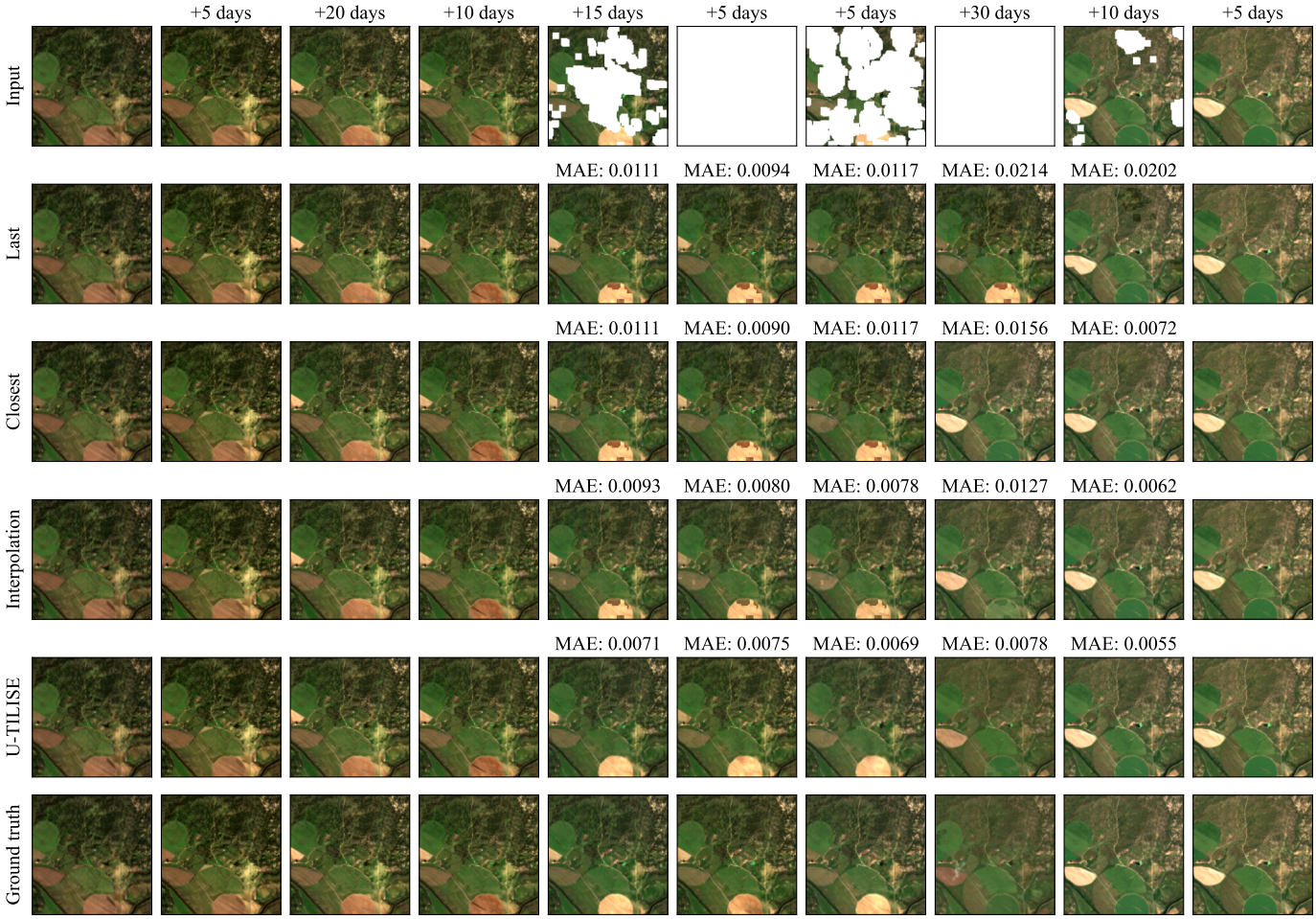

Fig. 2. Visual comparison of U-TILISE with selected baselines for a Sentinel-2 time series of the EarthNet2021 *iid* test split with predominantly gradually changing land cover. We show the true-color RGB composite for every image in the time series. The number above an image denotes its temporal distance to the previous observation in the sequence or its MAE, evaluated over all pixels that have been masked in the corresponding input image and across all spectral bands (R, G, B, NIR).

PSNR of almost 37 dB and a SSIM of 0.96, this still amounts to high fidelity and visual quality.

B. Comparison to baselines

Table I compares U-TILISE against the three baselines. We first discuss the performance on *iid* test sequences, corresponding to the imputation of new sequences acquired at previously seen locations. As expected, the *last* baseline, which copies the last valid observation, yields the largest reconstruction errors. Instead, cloning the temporally *closest* observation improves the reconstruction quality and the spectral fidelity markedly, reducing MAE, RMSE, and SAM by more than 10% and increasing PSNR by about 1 dB; mostly because the *last* heuristic degrades for longer data gaps where the same location is repeatedly occluded. *Linear interpolation* between the most recent and the next available observation

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2023.3333391

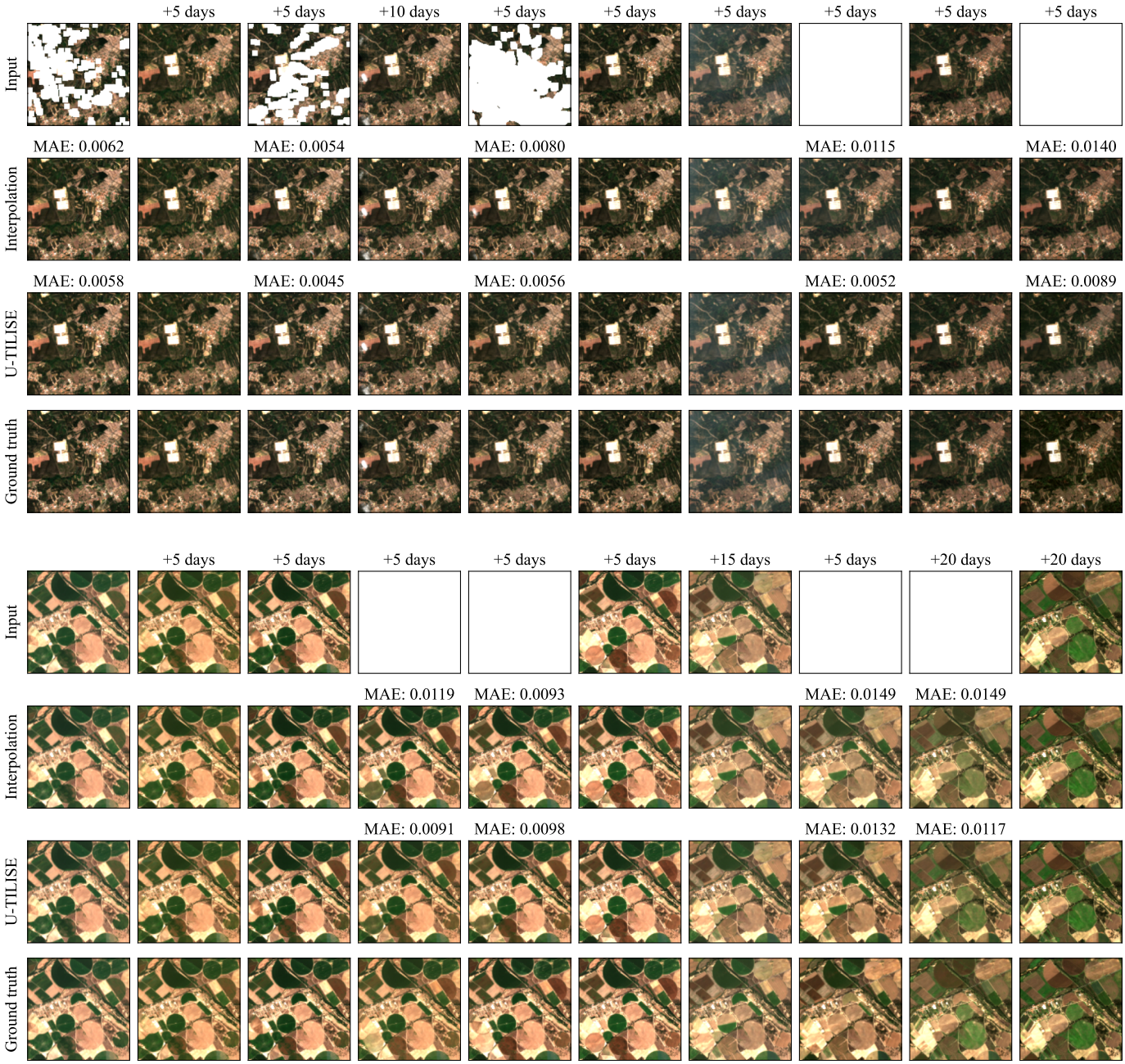

Fig. 3. Visual comparison (true-color RGB composites) of two Sentinel-2 time series of the EarthNet2021 *iid* test split, gap-filled using either U-TILISE or linear interpolation over time. Rows 1–4 depict a static scene, while rows 5–8 show a scene with sudden land cover changes due to agricultural activities by humans. The number above an image indicates its temporal distance to the previous observation in the sequence or its MAE, evaluated over all pixels that have been masked in the corresponding input image and across all spectral bands (R, G, B, NIR).

brings further gains. MAE, RMSE, and SAM drop by another $\approx 14\%$, and PSNR improves by 1.5 dB to 36.0 dB. U-TILISE consistently outperforms all baselines by a significant margin. Compared to linear interpolation, the MAE, RMSE, and SAM values decrease by 20%, and the PSNR increases by more than 1.5 dB. We observe similar trends when evaluating the *ood* test set of previously unseen locations. We note in passing that the *ood* test set is objectively more difficult: all methods perform slightly worse on it, although the baselines do not involve any learning and can, by definition, not overfit to specific geographic locations.

U-TILISE predicts a reflectance value for *every* spatio-temporal location in the output sequence, including pixels with valid input observations. In principle, those predicted values could deviate from the actual, observed values at cloud-free pixels.[6] For a complete evaluation, we thus also evaluate the spectral fidelity at pixels with valid input reflectances (Table I, columns 8–9). By construction, all baselines achieve the same, maximal performance, since they do not alter valid reflectance

[6]Obviously, one could, in practice, copy valid input pixels to the output. We regard this as a post-processing option available to any cloud removal method, not as a part of the actual model.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2023.3333391

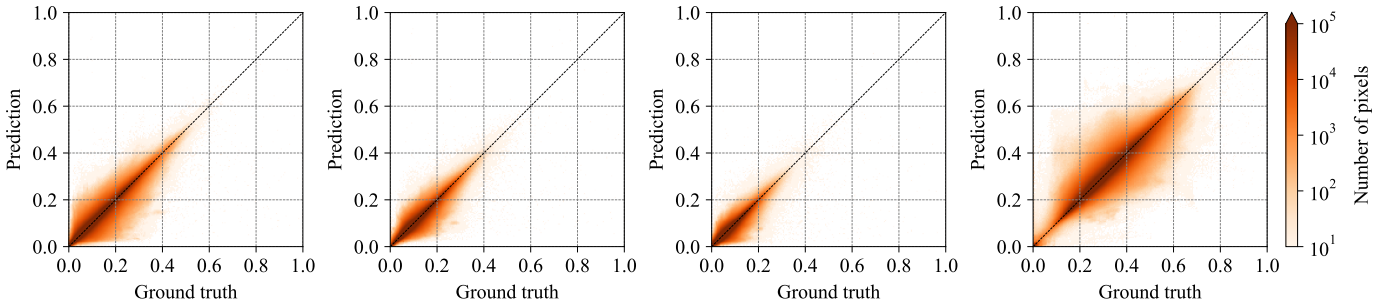

Fig. 4. Channel-wise ground truth reflectances vs. reflectances regressed by U-TILISE for missing pixels in optical time series of the EarthNet2021 dataset (*iid* split). From left to right: B4 (red), B3 (green), B2 (blue), and B8 (near-infrared).

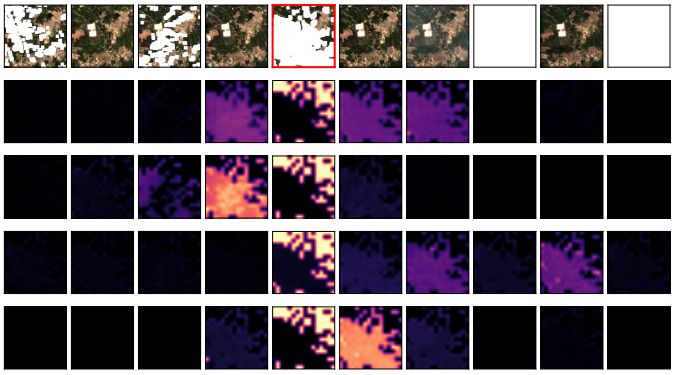

Fig. 5. Self-attention in the temporal encoder. We show the attention scores for imputing the 5th image in the example time series (highlighted with a red frame in row 1), displayed separately for each of the four attention heads (rows 2–5). The attention masks are bilinearly upsampled to the native spatial resolution of the input time series and color-coded from black (no attention) to yellow (maximum attention).

values. U-TILISE, on the other hand, must learn to preserve the reflectances at unoccluded (spatio-temporal) locations. It does that astonishingly well, with a MAE around $\frac{1}{5000}$ of the intensity range—less than the radiometric sensitivity of Sentinel-2.[7]

The estimates of U-TILISE are also qualitatively superior to those of the baselines, especially in the presence of significant spectral changes in time (cf. Fig. 2) and of remaining atmospheric effects (cf. Fig. 3, rows 1–4). A striking example is the southernmost, circular field in Fig. 2, for which U-TILISE smoothly transitions from dark to bright brown and then changes abruptly to dark green, which agrees well with the true evolution of the depicted scene. In contrast, copying or simple interpolation lead to evident visual artifacts.

C. Learned multi-temporal attention

The attention masks of the temporal self-attention mechanism encode the contribution of each input pixel to the regression of the output pixels.[8] This makes them a useful visual cue to determine on which input frames the learned model bases its predictions. Fig. 5 and Fig. 6 depict the attention masks for the first time series in Fig 3, demonstrating that U-TILISE indeed discovers, in a data-driven manner, which observations are most useful for its task. Not only is the attention low in data gaps (cf. Fig. 6), the model has also learned to preferentially attend to temporally close observations; to use information from unoccluded regions of the current frame, presumably to match its radiometry; and to let the attention heads specialize on different portions of the sequence (cf. Fig. 5).

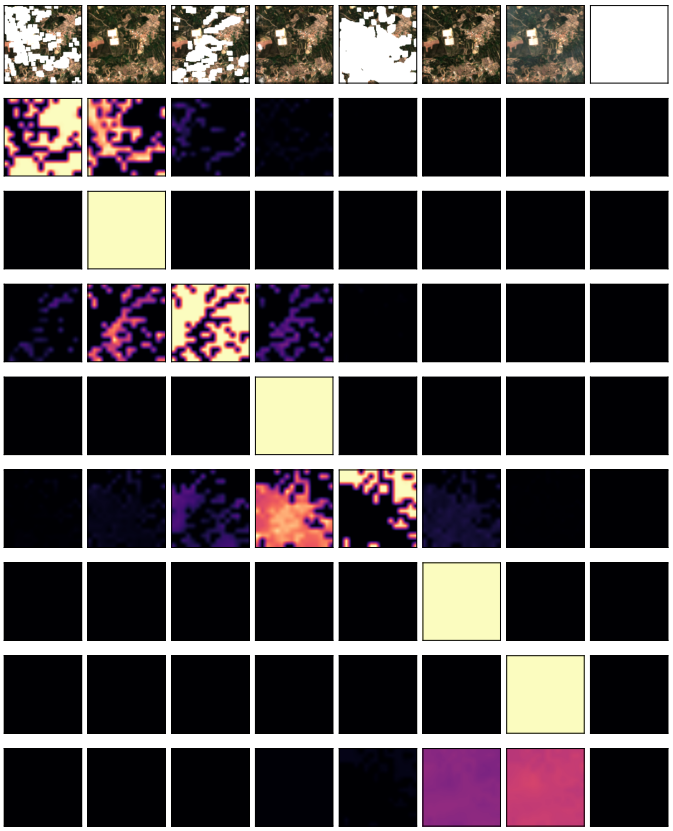

Fig. 6. Self-attention in the temporal encoder. For the time series in row 1. Rows 2–9 show the attention masks for one of the four heads, with rows corresponding to frames of the output (in temporal order). The masks are bilinearly upsampled to the native spatial resolution of the input and color-coded from black (no attention) to yellow (maximum attention). Note how the attention progressively moves through time to focus on unoccluded inputs and unoccluded pixels of partially occluded inputs, while it borrows information from temporally nearby frames where needed.

[7] https://sentinels.copernicus.eu/web/sentinel/user-guides/sentinel-2-msi/resolutions/radiometric

[8] Strictly speaking, the attention scores express the contribution w.r.t. the high-dimensional, but spatially coarsened latent embedding.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2023.3333391

TABLE II
QUANTITATIVE RESULTS OF DIFFERENT U-TILISE VARIANTS. WE ABLATE THE INFLUENCE OF THE TEMPORAL ENCODER AND TEMPORALLY WEIGHTED SKIP CONNECTIONS IN THE SPATIAL U-NET (ROWS 2–3), THE STRATEGY USED FOR POSITIONAL ENCODING (ROWS 4–6), THE NUMBER OF IMAGES T PROCESSED IN ONE SHOT (ROWS 7–8), AND THE NUMBER OF ATTENTION HEADS G OF THE TEMPORAL ENCODER (ROWS 9–11). THE METRICS ARE COMPUTED OVER ALL PIXELS (OR IMAGES, IN THE CASE OF SSIM) WITH MISSING DATA IN THE INPUT TIME SERIES.

Temp. attention	Weighted skip	Positional encoding	T	G	↓ MAE [ρ_{TOA}]	↓ RMSE [ρ_{TOA}]	↓ SAM [°]	↑ PSNR [dB]	↑ SSIM [-]
✓	✓	day within year	10	4	0.0100	0.0160	2.16	36.82	0.964
✓		day within year	10	4	0.0120	0.0187	2.75	35.14	0.936
		day within year	10	4	0.0339	0.0498	8.57	26.49	0.806
✓	✓	day within sequence	10	4	0.0098	0.0158	2.07	37.00	**0.965**
✓	✓	enumeration	10	4	0.0124	0.0199	2.61	34.93	0.953
✓	✓	-	10	4	0.0142	0.0228	3.14	33.91	0.947
✓	✓	day within year	5	4	0.0102	0.0163	2.18	36.66	0.963
✓	✓	day within year	15	4	0.0099	0.0159	2.14	36.89	0.964
✓	✓	day within year	10	2	0.0101	0.0162	2.18	36.77	0.963
✓	✓	day within year	10	8	0.0098	0.0158	2.12	36.99	0.964
✓	✓	day within year	10	16	**0.0097**	**0.0156**	**2.10**	**37.09**	**0.965**

D. Importance of temporal encoding

We go on to study the influence of different network configurations, starting with the temporal encoder. To this end, we conduct two ablation experiments: *(i)* we replace the temporally weighted skip connections between corresponding layers of the spatial encoder and decoder with ordinary skip connections, and *(ii)* we remove the temporal encoder altogether, resulting in a standard U-Net architecture that processes each frame of the input sequence independently. As expected, we observe a significant deterioration in all evaluation metrics when U-TILISE cannot exploit the time series to fill in missing image content (Table II, 3rd row): without a mechanism to represent spatio-temporal relations, the model must hallucinate the missing spectral values only from the unoccluded context in the same image. In the extreme case of a fully occluded view, this means blindly synthesizing a multi-spectral image from a blank slate. Due to this severe ill-conditioning, the model merely learns to reproduce observed values while replacing missing pixels with the average reflectance value of the training set. On the contrary, as soon as the temporal encoder is added in the bottleneck of the U-Net, while still using ordinary skip connections otherwise, U-TILISE learns to leverage the temporal context to inform the reconstruction of missing pixels. MAE, RMSE, and SAM are about 3× lower than without temporal encoding, while the PSNR increases by 8.7 dB (Table II, 2nd row). Visually, U-TILISE predicts plausible spectral intensities and colors in occluded image regions but fails to recover fine-grained texture details (Fig. 7(a)). The full model, which also uses the upsampled attention maps in the skip connections, manages to restore visibly more high-frequency details (Fig. 7(b)). The relative improvement amounts to ≈ 15% in MAE and RMSE, and more than 20% in terms of SAM (Table II, 1st row). PSNR improves by another 1.7 dB.

E. Choice of positional encoding

Next, we evaluate the influence of the positional encoding scheme (cf. Section III-D) on the model output. We found experimentally that information about the image order and the relative temporal distance between observations is crucial to reconstruct realistic, gap-free time series. Without any positional encoding, the performance of U-TILISE drops significantly; MAE, RMSE, and SAM increase by ≈ 40% and PSNR drops by almost 3 dB (Table II, 6th row). Injecting ordinal information improves MAE and RMSE by 13% and SAM by 17%, while also improving the perceptual similarity to the target sequence (Table II, 5th row). Encoding the temporal offset from the first observation in the sequence rather than the ordinal position brings another improvement of ≈ 20% in MAE, RMSE, and SAM and boosts PSNR by 2 dB (Table II, 4th row). We find only a tiny difference between encoding the temporal distance to the first observation or encoding the acquisition date of every observation relative to the 1st January of the respective calendar year. We speculate that the latter strategy suffers from a bias inherent in our training data: most of the EarthNet2021 time series are captured between May and October; likely, the dataset does not offer sufficient variability to extract an expressive prior over seasonal patterns.

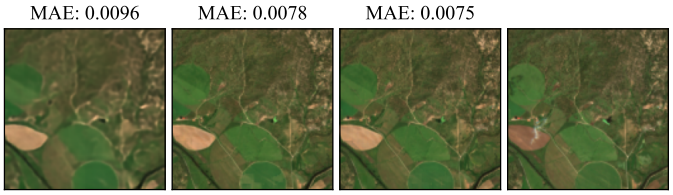

Fig. 7. Visual comparison of different U-TILISE variants for the 8th time step in the sequence from Fig. 2 (the 2nd totally masked image). (a) U-TILISE with ordinary skip connections; (b) U-TILISE with default parameter settings; (c) U-TILISE with 16 attention heads; (d) ground truth.

F. Influence of input sequence length T

To test the sensitivity of U-TILISE to the number of images processed as one sequence, we define two model variants by varying the temporal window length T. In detail, we retrain U-TILISE once with training sequences that are randomly trimmed to a maximum temporal length of five images ($T = 5$), and once with time series that comprise up to 15 images ($T = 15$). At test time, we always process the

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2023.3333391

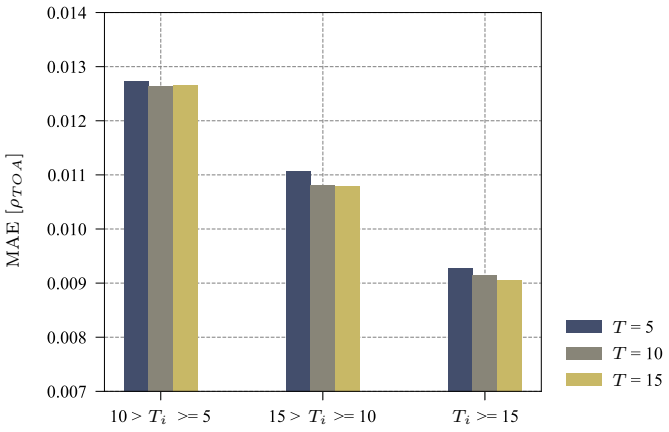

Fig. 8. MAE as a function of the full sequence length T_i and the temporal window T of U-TILISE.

full-length time series, employing a sliding window scheme if the temporal length T_i is larger than the maximum length T used during training.

We find that all evaluation metrics remain relatively stable when varying the temporal window of U-TILISE (Table II, rows 7–8), with fluctuations below 2% compared to our default setting ($T = 10$). For a more fine-grained analysis, we thus separately measure the performance for time series consisting of (*i*) at most 9 images, (*ii*) 10 to 14 images, or (*iii*) 15 or more images. As shown in Fig. 8, the MAE at imputed pixels slightly reduces with longer time series and temporal window T, indicating that U-TILISE can exploit long temporal context if needed.

G. Number of attention heads

U-TILISE is fairly robust with varying numbers of temporal attention heads. We find marginal quantitative gains when adding more heads (Table II, rows 9-11), and also only small differences in visual quality (Fig. 7(c)).

H. Time series with real data gaps

In the last experiment, we use the trained U-TILISE model, unaltered, to impute missing pixels in time series with real data gaps. This scenario corresponds to the practical application case, where the masked pixels are truly unobserved. Of course, this also implies that the resulting outputs can only be assessed through visual inspection, since no ground truth exists for the masked areas. Note that the application to real cloudy time series may, from a machine learning perspective, involve some degree of generalization. As clouds do not have sharp boundaries, the unmasked regions just outside the cloud mask may, in some cases, still be affected by thin clouds and haze. During training, where cloud-free images are synthetically masked, the model has not been exposed to such a situation.

Fig. 9 depicts imputation results for two representative 30-frame time series from the EarthNet2021 test set. The original, observed time series suffer from severe data gaps from clouds, shadows, haze, and missing images. Furthermore, the second example exhibits a rather long period without any valid observations. Despite these challenges, U-TILISE creates realistic, gap-free time series of high visual quality. The reconstructed time series do occasionally suffer from remaining clouds or haze, if they were missed by the cloud masking algorithm (Fig. 9, 5th last frame of the second example).

VII. Conclusion

We have presented U-TILISE, a learned sequence-to-sequence model for data imputation in optical satellite image time series. The model combines 2D convolutions over the spatial and spectral dimensions and 1D self-attention across time into an efficient prior over multi-spectral and multi-temporal reflectance patterns. Given an optical time series in which invalid reflectance values are masked, U-TILISE creates a coherent time series with a clean, cloud-free image at every time step of the input.

In a series of experiments, we have shown that U-TILISE reconstructs gap-free Sentinel-2 sequences with high accuracy. It removes clouds and cloud shadows of various shapes and sizes, manages to recover multiple consecutive frames of missing data, and generalizes to previously unseen geographical locations. On the EarthNet2021 dataset, the average MAE within the data gaps is in the order of 1% of the intensity range, and the PSNR is ≈ 38 dB.

A limitation of the current approach is that it relies on cloud (and cloud shadow) masks as auxiliary input. Detecting cloudy pixels in remotely sensed imagery is challenging and still not completely solved. Mistakes of the preceding cloud detector limit the performance of U-TILISE, since it implicitly learns to preserve input reflectances that are valid according to the masks. Furthermore, the model may of course reconstruct radiometrically plausible but incorrect transitions in cases where the available information is too sparse to determine when a sudden change has occurred, such as, for instance, a harvesting event during a multi-frame data gap.

An interesting future research direction is to eliminate the need for external cloud masks and instead design the model such that it implicitly performs the detection of invalid pixels. Another useful extension for users of the U-TILISE output will be to integrate a probabilistic deep learning scheme and supplement the output with spatially and temporally resolved uncertainty estimates.

Acknowledgment

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via Contract #2021-21040700001. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2023.3333391

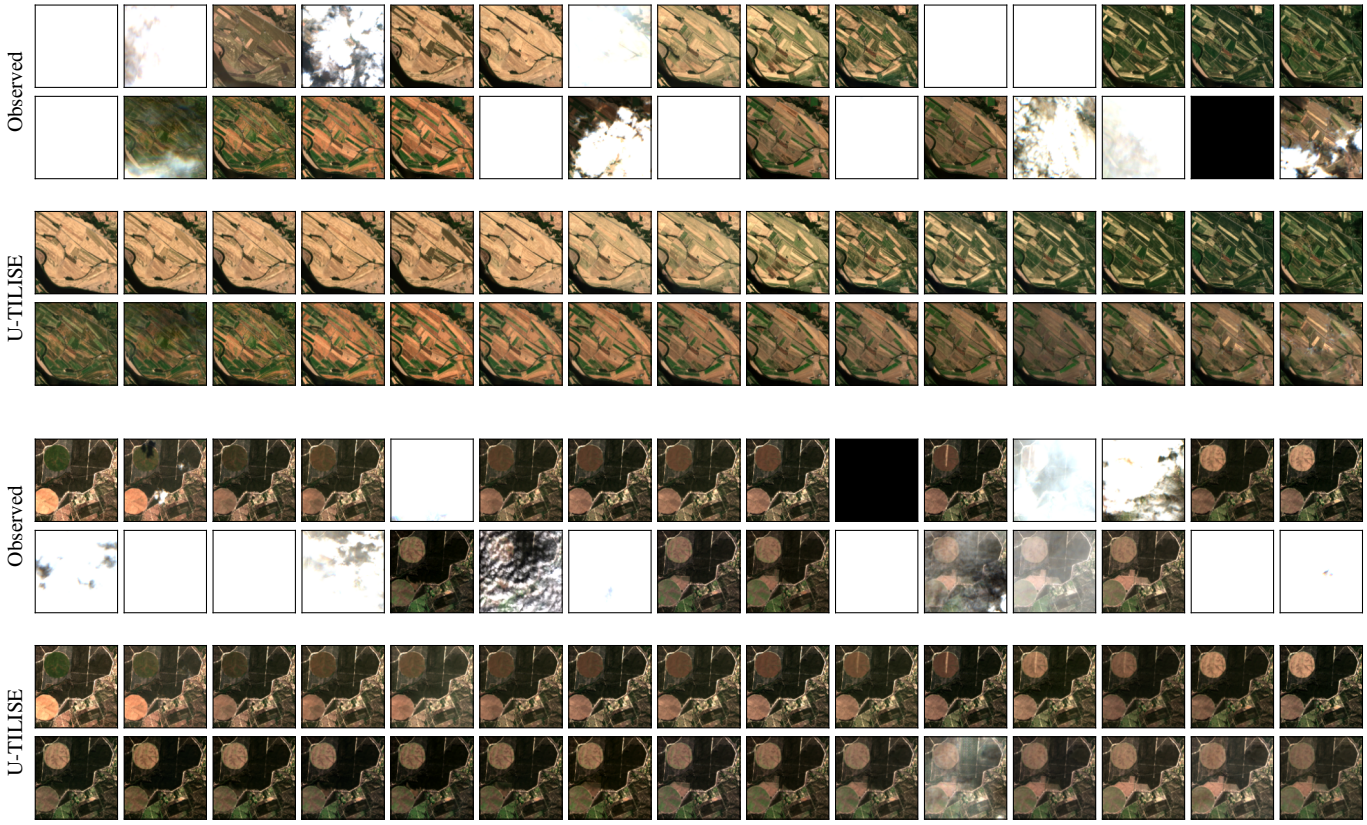

Fig. 9. Imputation results for two time series of the EarthNet2021 dataset with real data gaps due to clouds, cloud shadows, and missing frames (shown in black). Example 1 is from the *ood* test split and example 2 from the *iid* test split.

APPENDIX
COMPLEMENTARY EXPERIMENTS

Some authors [31], [61] have advocated the use of co-registered SAR observations, which are largely unaffected by the atmosphere, to support optical image imputation. It is a natural idea to also augment U-TILISE with a time series of spatially and temporally co-registered SAR images. Technically, this is straightforward: we add the 2-channel SAR images (ortho-rectified VV/VH log-amplitude) as additional input channels that need not be reconstructed and, accordingly, increase the filter depth of the first layer by two. We train and test this multi-modal variant of U-TILISE in a simulated setting using the SEN12MS-CR-TS dataset [31], a multi-modal and multi-temporal dataset specifically designed for multi-modal cloud removal. Unfortunately, it turns out that the dataset is not only considerably smaller than EarthNet2021 (see Table III), but its sequences also exhibit comparatively lower temporal variability and dynamics and do not allow a conclusive comparison. We refrain from drawing firm conclusions about the impact of complementary SAR observations and instead conclude that a more informative dataset must be created to investigate the issue, possibly by augmenting EarthNet2021 with SAR observations.

SEN12MS-CR-TS dataset

SEN12MS-CR-TS [31] comprises about 15 000 globally sampled Sentinel-2 time series (Level-1C with top-of-atmosphere reflectances) from 2018 with a spatial extent of 256×256 pixels (2.56×2.56 km in scene space). Each time series contains 30 images, with varying temporal spacing of 5 to 15 days between consecutive observations. The images encompass all 13 spectral bands, upsampled to 10 m resolution. Furthermore, every optical image is paired with a spatially co-registered, temporally close (but not synchronous) C-band SAR image with two channels representing the σ_0 backscatter coefficients in the VV and VH polarizations, in units of decibels (dB). The temporal offset between paired optical and SAR observations is three days on average, although, for 5.5% of the pairs, the temporal difference is over one week. The dataset also includes pixel-wise cloud probabilities and binary cloud masks, produced with the `S2Cloudless` detector [76].

Experimental setup

We adopt the preprocessing protocol of [31], [32] and value-clip the optical images to the range $[0, 10\,000]$ and the SAR images to $[-25, 0]$, followed by normalization to the unit range $[0, 1]$. Like for EarthNet2021, we extract cloud-free optical time series for training and evaluation (cf. Section IV-B). In rare cases, SEN12MS-CR-TS sequences exhibit data gaps of several consecutive months. To limit potential land cover and seasonal changes to a reasonable range, we temporally trim the time series such that the temporal spacing between adjacent valid frames is at most four weeks. The resulting time series are, on average, shorter than those of EarthNet2021, and they

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2023.3333391

TABLE III
ACQUISITION DETAILS OF THE TWO SENTINEL-2 DATASETS. WE LIST THE NUMBER OF SEQUENCES, THEIR AVERAGE LENGTH, AND THE TEMPORAL RESOLUTION; SEPARATELY FOR THE TRAINING, VALIDATION, AND TEST PARTS (FOR EARTHNET2021, FURTHER DIVIDED INTO *iid* AND *ood* TEST SETS). WE ALSO SHOW THE STATISTICS FOR THE SIMULATED TRAINING SEQUENCES, AFTER REMOVING IMAGES WITH ACTUAL CLOUDS.

	EarthNet2021 [37]	SEN12MS-CR-TS [31]
Acquisition		
Spatial coverage	Europe	global
Time period	Nov 2016–May 2020	2018
Spectral bands	RGB, NIR	all 13 bands
Image size [pix]	128×128	256×256
Spatial resolution [m]	20	10
Temporal length [-]	30	30
Temporal resolution [days]	regular, 5	irregular, avg. 12
No. of sequences [-]		
train	18 955	10 167
val	4 949	1 410
test	4 219 | 4 214	3 716
Cloud-free subsequences		
No. of sequences [-]		
train	18 823	5 618
val	4 944	937
test	3 948 | 4 116	2 238
Temporal length [-]		
train	15.1 ± 4.5	11.1 ± 6.0
val	16.4 ± 5.2	7.2 ± 2.1
test	13.3 ± 5.0 | 13.8 ± 5.5	8.8 ± 3.7
Temporal resolution [days]		
train	10.2 ± 3.3	14.9 ± 2.3
val	9.6 ± 3.4	14.4 ± 1.9
test	11.3 ± 4.6 | 11.2 ± 4.6	16.1 ± 3.1

have about 50% larger temporal spacing between consecutive frames (cf. Table III).

After removing real data gaps, we introduce synthetic gaps into the optical images (cf. Section IV-C) and concatenate the resulting, masked optical time series with the (unmodified) SAR time series along the channel dimension to produce multi-modal input for U-TILISE.

Training details

We use the Adam optimizer [81] with hyper-parameters $\{\beta_1$=0.9, β_2=0.999$\}$, batch size 3, and a weight decay of 10^{-5}. The base learning rate of $2 \cdot 10^{-4}$ is reduced by 50% every 80 training epochs. Due to the larger spatial dimensions of the input time series (256×256 pixels, compared to 128×128 pixels in EarthNet2021), we add an additional convolutional block in the spatial encoder and decoder, such that the (spatial) dimension of 16×16 pixels in the bottleneck is the same as for EarthNet2021.

Results

We provide quantitative results in Table IV and a visual example in Fig. 10. As already observed with EarthNet2021, linear interpolation between the most recent and the next available observation performs significantly better than the more widely used baselines that replicate either the *last* or the temporally *closest* observation. U-TILISE achieves marginally better error metrics with SAR guidance than without, but the differences (<0.05% of the intensity range in MAE and RMSE, $<0.2^\circ$ in SAM, <0.2% in SSIM) are negligible and well within the stochastic fluctuations of deep network training. Moreover, also the linear interpolation baseline is on par with both variants, and all three results remain well below the fidelity achieved on EarthNet2021 (Table I), although the numbers are not directly comparable since SEN12MS-CR-TS includes 13 bands of which 9 have been *up*sampled to 10 m GSD, whereas EarthNet2021 consists of 4 bands that were *down*sampled to 20 m GSD. Upon inspection, we find that SEN12MS-CR-TS contains many sequences where the land cover is static and largely homogeneous. The results we obtain on SEN12MS-CR-TS neither confirm nor rule out a possible benefit through SAR guidance. We believe that a larger dataset with more non-linear temporal variations will be needed to carry out a conclusive comparison.

REFERENCES

[1] M. Rußwurm and M. Körner, "Self-attention for raw optical satellite time series classification," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 169, pp. 421–435, 2020.

[2] V. S. F. Garnot, L. Landrieu, S. Giordano, and N. Chehata, "Satellite image time series classification with pixel-set encoders and temporal self-attention," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 12 325–12 334.

[3] D. Hong, Z. Han, J. Yao, L. Gao, B. Zhang, A. Plaza, and J. Chanussot, "SpectralFormer: Rethinking hyperspectral image classification with transformers," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2021.

[4] V. S. F. Garnot and L. Landrieu, "Panoptic segmentation of satellite image time series with convolutional temporal attention networks," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 4872–4881.

[5] M. Tarasiou, E. Chavez, and S. Zafeiriou, "ViTs for SITS: Vision transformers for satellite image time series," *arXiv preprint arXiv:2301.04944*, 2023.

[6] P. Potapov, X. Li, A. Hernandez-Serna, A. Tyukavina, M. C. Hansen, A. Kommareddy, A. Pickens, S. Turubanova, H. Tang, C. E. Silva *et al.*, "Mapping global forest canopy height through integration of GEDI and Landsat data," *Remote Sensing of Environment*, vol. 253, p. 112165, 2021.

[7] A. Becker, S. Russo, S. Puliti, N. Lang, K. Schindler, and J. D. Wegner, "Country-wide retrieval of forest structure from optical and SAR satellite imagery with deep ensembles," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 195, pp. 269–286, 2023.

[8] R. C. Daudt, H. Wulf, E. D. Hafner, Y. Bühler, K. Schindler, and J. D. Wegner, "Snow depth estimation at country-scale with high spatial and temporal resolution," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 197, pp. 105–121, 2023.

[9] X.-P. Song, M. C. Hansen, S. V. Stehman, P. V. Potapov, A. Tyukavina, E. F. Vermote, and J. R. Townshend, "Global land change from 1982 to 2016," *Nature*, vol. 560, no. 7720, pp. 639–643, 2018.

[10] I. K. Wernick, P. Ciais, J. Fridman, P. Högberg, K. T. Korhonen, A. Nordin, and P. E. Kauppi, "Quantifying forest change in the European Union," 2021.

[11] N. Lang, W. Jetz, K. Schindler, and J. D. Wegner, "A high-resolution canopy height model of the Earth," *arXiv preprint arXiv:2204.08322*, 2022.

[12] D. J. Mulla, "Twenty five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps," *Biosystems Engineering*, vol. 114, no. 4, pp. 358–371, 2013.

[13] T. B. Hank, K. Berger, H. Bach, J. G. Clevers, A. Gitelson, P. Zarco-Tejada, and W. Mauser, "Spaceborne imaging spectroscopy for sustainable agriculture: Contributions and challenges," *Surveys in Geophysics*, vol. 40, pp. 515–551, 2019.

[14] J. Segarra, M. L. Buchaillot, J. L. Araus, and S. C. Kefauver, "Remote sensing for precision agriculture: Sentinel-2 improved features and applications," *Agronomy*, vol. 10, no. 5, p. 641, 2020.

[15] D. K. Bolton, J. M. Gray, E. K. Melaas, M. Moon, L. Eklundh, and M. A. Friedl, "Continental-scale land surface phenology from harmonized Landsat 8 and Sentinel-2 imagery," *Remote Sensing of Environment*, vol. 240, p. 111685, 2020.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2023.3333391

TABLE IV
QUANTITATIVE COMPARISON OF U-TILISE WITH CONVENTIONAL BASELINES FOR THE SEN12MS-CR-TS DATASET. THE METRICS ARE COMPUTED OVER ALL PIXELS (OR IMAGES, IN THE CASE OF SSIM) WITH MISSING DATA IN THE INPUT TIME SERIES. WE TRAIN AND EVALUATE U-TILISE ONCE ONLY WITH OPTICAL TIME SERIES AND ONCE WITH ADDITIONAL SAR INPUT.

Method	↓ MAE [ρ_{TOA}]	↓ RMSE [ρ_{TOA}]	↓ SAM [°]	↑ PSNR [dB]	↑ SSIM [-]
Last	0.0178	0.0268	5.06	32.18	0.945
Closest	0.0165	0.0250	4.77	32.76	0.949
Linear interpolation	0.0144	0.0218	4.22	34.02	0.957
U-TILISE, w/o SAR (ours)	0.0149	0.0222	4.39	33.77	0.955
U-TILISE, w/ SAR (ours)	0.0146	0.0217	4.32	33.94	0.956

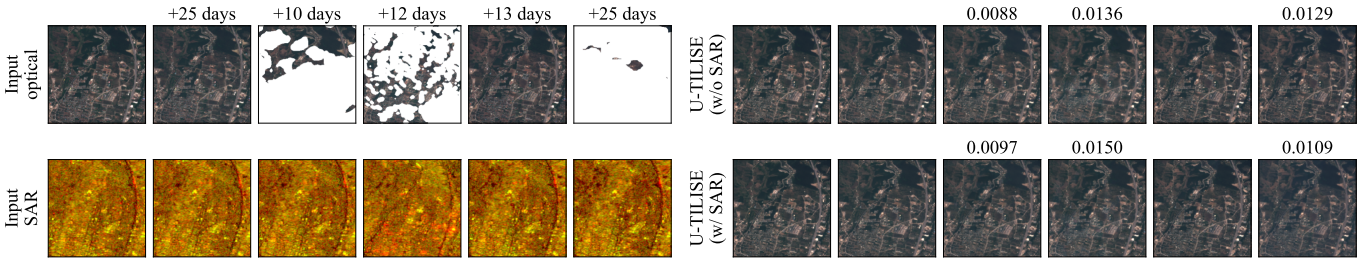

Fig. 10. Visual comparison of two U-TILISE variants on the SEN12MS-CR-TS dataset. On the left are the inputs for an exemplary sequence. For multi-spectral data, the RGB channels are displayed; SAR data are rendered as 2-channel image composites (VV/VH amplitude). Numbers indicate the temporal spacing from the preceding image. U-TILISE prediction from only optical time series (top) and from combined optical and SAR input (bottom) are visually indistinguishable. Numbers are mean absolute errors over all masked pixels, across all 13 spectral bands.

[16] J. W. Rouse, R. H. Haas, J. A. Schell, D. W. Deering *et al.*, "Monitoring vegetation systems in the great plains with ERTS," *NASA Spec. Publ*, vol. 351, no. 1, p. 309, 1974.

[17] G. Misra, F. Cawkwell, and A. Wingler, "Status of phenological research using Sentinel-2 data: A review," *Remote Sensing*, vol. 12, no. 17, p. 2760, 2020.

[18] A. Garioud, S. Valero, S. Giordano, and C. Mallet, "Recurrent-based regression of Sentinel time series for continuous vegetation monitoring," *Remote Sensing of Environment*, vol. 263, p. 112419, 2021.

[19] M. D. King, S. Platnick, W. P. Menzel, S. A. Ackerman, and P. A. Hubanks, "Spatial and temporal distribution of clouds observed by MODIS onboard the terra and aqua satellites," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 7, pp. 3826–3852, 2013.

[20] J. Gawlikowski, P. Ebel, M. Schmitt, and X. X. Zhu, "Explaining the effects of clouds on remote sensing scene classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 9976–9986, 2022.

[21] Z. Gu, P. Ebel, Q. Yuan, M. Schmitt, and X. X. Zhu, "Explicit haze & cloud removal for global land cover classification," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2022, pp. 1–6.

[22] K. Enomoto, K. Sakurada, W. Wang, H. Fukui, M. Matsuoka, R. Nakamura, and N. Kawaguchi, "Filmy cloud removal on satellite imagery with multispectral conditional generative adversarial nets," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2017, pp. 48–56.

[23] C. Grohnfeldt, M. Schmitt, and X. Zhu, "A conditional generative adversarial network to fuse SAR and multispectral optical data for cloud removal from sentinel-2 images," in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2018, pp. 1726–1729.

[24] P. Singh and N. Komodakis, "Cloud-gan: Cloud removal for sentinel-2 imagery using a cyclic consistent generative adversarial networks," in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2018, pp. 1772–1775.

[25] J. Bermudez, P. Happ, D. Oliveira, and R. Feitosa, "SAR to optical image synthesis for cloud removal with generative adversarial networks," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 4, no. 1, 2018.

[26] J. D. Bermudez, P. N. Happ, R. Q. Feitosa, and D. A. Oliveira, "Synthesis of multispectral optical images from SAR/optical multitemporal data using conditional generative adversarial networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 8, pp. 1220–1224, 2019.

[27] L. Wang, X. Xu, Y. Yu, R. Yang, R. Gui, Z. Xu, and F. Pu, "SAR-to-optical image translation using supervised cycle-consistent adversarial networks," *IEEE Access*, vol. 7, pp. 129 136–129 149, 2019.

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

[29] J. Zhang, J. Zhou, and X. Lu, "Feature-guided SAR-to-optical image translation," *IEEE Access*, vol. 8, pp. 70 925–70 937, 2020.

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

[31] P. Ebel, Y. Xu, M. Schmitt, and X. X. Zhu, "SEN12MS-CR-TS: A remote-sensing data set for multimodal multitemporal cloud removal," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022.

[32] F. Xu, Y. Shi, P. Ebel, L. Yu, G.-S. Xia, W. Yang, and X. X. Zhu, "GLF-CR: SAR-enhanced cloud removal with global–local fusion," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 192, pp. 268–278, 2022.

[33] V. Sarukkai, A. Jain, B. Uzkent, and S. Ermon, "Cloud removal in satellite images using spatiotemporal generative networks," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020, pp. 1785–1794.

[34] Q. Zhang, Q. Yuan, J. Li, Z. Li, H. Shen, and L. Zhang, "Thick cloud and cloud shadow removal in multitemporal imagery using progressively spatio-temporal patch group deep learning," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 162, pp. 148–160, 2020.

[35] S. Oehmcke, T.-H. K. Chen, A. V. Prishchepov, and F. Gieseke, "Creating cloud-free satellite imagery from image time series with deep learning," in *ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data*, 2020, pp. 1–10.

[36] Y. Chen, Q. Weng, L. Tang, X. Zhang, M. Bilal, and Q. Li, "Thick clouds removing from multitemporal Landsat images using spatiotemporal neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2020.

[37] C. Requena-Mesa, V. Benson, M. Reichstein, J. Runge, and J. Denzler, "EarthNet2021: A large-scale dataset and challenge for earth surface forecasting as a guided video prediction task." in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2021, pp. 1132–1142.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2023.3333391

[38] J. Long, Z. Shi, W. Tang, and C. Zhang, "Single remote sensing image dehazing," *IEEE Geoscience and Remote Sensing Letters*, vol. 11, no. 1, pp. 59–63, 2013.

[39] M. Xu, M. Pickering, A. J. Plaza, and X. Jia, "Thin cloud removal based on signal transmission principles and spectral mixture analysis," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 3, pp. 1659–1669, 2015.

[40] H. Lv, Y. Wang, and Y. Shen, "An empirical and radiative transfer model based algorithm to remove thin clouds in visible bands," *Remote Sensing of Environment*, vol. 179, pp. 183–195, 2016.

[41] H. Shen, H. Li, Y. Qian, L. Zhang, and Q. Yuan, "An effective thin cloud removal procedure for visible remote sensing images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 96, pp. 224–235, 2014.

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

[43] M. K.-P. Ng, Q. Yuan, L. Yan, and J. Sun, "An adaptive weighted tensor completion method for the recovery of remote sensing images with missing data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 6, pp. 3367–3381, 2017.

[44] X. Li, L. Wang, Q. Cheng, P. Wu, W. Gan, and L. Fang, "Cloud removal in remote sensing images using nonnegative matrix factorization and error correction," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 148, pp. 103–113, 2019.

[45] J. Lin, T.-Z. Huang, X.-L. Zhao, Y. Chen, Q. Zhang, and Q. Yuan, "Robust thick cloud removal for multitemporal remote sensing images using coupled tensor factorization," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2022.

[46] E. H. Helmer and B. Ruefenacht, "Cloud-free satellite image mosaics with regression trees and histogram matching." *Photogrammetric Engineering and Remote Sensing*, vol. 71, pp. 1079–1089, 2005.

[47] D.-C. Tseng, H.-T. Tseng, and C.-L. Chien, "Automatic cloud removal from multi-temporal SPOT images," *Applied Mathematics and Computation*, vol. 205, no. 2, pp. 584–600, 2008.

[48] C.-H. Lin, P.-H. Tsai, K.-H. Lai, and J.-Y. Chen, "Cloud removal from multitemporal satellite images using information cloning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 1, pp. 232–241, 2012.

[49] C.-H. Lin, K.-H. Lai, Z.-B. Chen, and J.-Y. Chen, "Patch-based information reconstruction of cloud-contaminated multitemporal images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 1, pp. 163–174, 2013.

[50] H. Shen and L. Zhang, "A map-based algorithm for destriping and inpainting of remotely sensed images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 5, pp. 1492–1502, 2008.

[51] A. Maalouf, P. Carré, B. Augereau, and C. Fernandez-Maloigne, "A bandelet-based inpainting technique for clouds removal from remotely sensed images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 7, pp. 2363–2371, 2009.

[52] F. Meng, X. Yang, C. Zhou, and Z. Li, "A sparse dictionary learning-based adaptive patch inpainting method for thick clouds removal from high-spatial resolution remote sensing imagery," *Sensors*, vol. 17, no. 9, p. 2130, 2017.

[53] Y. Zeng, J. Fu, and H. Chao, "Learning joint spatial-temporal transformations for video inpainting," in *European Conference on Computer Vision (ECCV)*, 2020, pp. 528–543.

[54] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[55] K. Perlin, "Improving noise," in *ACM SIGGRAPH Conference on Computer Graphics and Interactive Techniques*, 2002, pp. 681–682.

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

[57] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017, pp. 2223–2232.

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

[59] H. Liu, B. Huang, and J. Cai, "Thick cloud removal under land cover changes using multisource satellite imagery and a spatiotemporal attention network," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[60] X. Yang, Y. Zhao, and R. R. Vatsavai, "Deep residual network with multi-image attention for imputing under clouds in satellite imagery," in *IEEE International Conference on Pattern Recognition (ICPR)*, 2022, pp. 643–649.

[61] A. Sebastianelli, E. Puglisi, M. P. Del Rosso, J. Mifdal, A. Nowakowski, P. P. Mathieu, F. Pirri, and S. L. Ullo, "PLFM: Pixel-level merging of intermediate feature maps by disentangling and fusing spatial and temporal data for cloud removal," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2022.

[62] P. Ebel, M. Schmitt, and X. X. Zhu, "Internal learning for sequence-to-sequence cloud removal via synthetic aperture radar prior information," in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2021, pp. 2691–2694.

[63] T. Peng, M. Liu, X. Liu, Q. Zhang, L. Wu, and X. Zou, "Reconstruction of optical image time series with unequal lengths SAR based on improved sequence–sequence model," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–17, 2022.

[64] M. Zhao, P. Olsen, and R. Chandra, "Seeing through clouds in satellite images," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[65] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[66] M. O. Turkoglu, S. D'Aronco, J. D. Wegner, and K. Schindler, "Gating revisited: Deep multi-layer RNNs that can be trained," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 8, pp. 4081–4092, 2021.

[67] H. Zhang, L. Mai, N. Xu, Z. Wang, J. Collomosse, and H. Jin, "An internal learning approach to video inpainting," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 2720–2729.

[68] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep image prior," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 9446–9454.

[69] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-assisted Intervention (MICCAI)*, 2015, pp. 234–241.

[70] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[71] R. Xiong, Y. Yang, D. He, K. Zheng, S. Zheng, C. Xing, H. Zhang, Y. Lan, L. Wang, and T. Liu, "On layer normalization in the transformer architecture," in *International Conference on Machine Learning*, 2020, pp. 10 524–10 533.

[72] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[73] Y. Wu and K. He, "Group normalization," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19.

[74] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," *arXiv preprint arXiv:1606.08415*, 2016.

[75] N. Metzger, M. O. Turkoglu, S. D'Aronco, J. D. Wegner, and K. Schindler, "Crop classification under varying cloud cover with neural ordinary differential equations," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2021.

[76] L. Baetens, C. Desjardins, and O. Hagolle, "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*, vol. 11, no. 4, p. 433, 2019.

[77] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 5505–5514.

[78] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, "Image inpainting for irregular holes using partial convolutions," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 85–100.

[79] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Free-form image inpainting with gated convolution," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 4471–4480.

[80] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019.

[81] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2023.3333391

[82] F. A. Kruse, A. Lefkoff, J. Boardman, K. Heidebrecht, A. Shapiro, P. Barloon, and A. Goetz, "The spectral image processing system (SIPS)—interactive visualization and analysis of imaging spectrometer data," *Remote Sensing of Environment*, vol. 44, no. 2-3, pp. 145–163, 1993.

[83] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[84] J. Inglada, M. Arias, B. Tardy, O. Hagolle, S. Valero, D. Morin, G. Dedieu, G. Sepulcre, S. Bontemps, P. Defourny *et al.*, "Assessment of an operational system for crop type map production using high temporal and spatial resolution satellite optical imagery," *Remote Sensing*, vol. 7, no. 9, pp. 12 356–12 379, 2015.

Corinne Stucker received the B.Sc. degree in Geomatics Engineering and Planning in 2015 and the M.Sc. degree in Geomatics Engineering in 2017 from ETH Zurich, Zurich, Switzerland. She completed her Ph.D. at ETH Zurich in July 2023 under the supervision of Prof. Dr. Konrad Schindler. Her research lies at the intersection of 3D computer vision, machine learning, and remote sensing, focusing on scene representations for dense 3D reconstruction from satellite imagery.

Vivien Sainte Fare Garnot is a PostDoc at the Institute for Computational Science at the University of Zurich, Zurich, Switzerland, and is affiliated with the EcoVision Lab since 2022. He received his Ph.D. degree in Geospatial Computer Vision in January 2022 at the French mapping agency (IGN, Gustave Eiffel University). His thesis explored novel deep learning architectures for crop type mapping from satellite image time series. Among other aspects, it focused on leveraging the temporal dimension of such data with attention mechanisms. The thesis also proposed a novel method to leverage the hierarchical structure of the crop type taxonomy and showed how to address crop type mapping as a panoptic segmentation task. For his PostDoc, he is involved in the development of machine learning methods for forest monitoring and other ecological applications. More broadly, Vivien is interested in methods that leverage remote sensing data to improve our understanding of the Earth's evolution.

Konrad Schindler (Senior Member, IEEE) received the Diplomingenieur (M.Tech.) degree from the Vienna University of Technology, Vienna, Austria, in 1999 and the Ph.D. degree from the Graz University of Technology, Graz, Austria, in 2003. He was a photogrammetric engineer in the private industry and held research positions at the Graz University of Technology; Monash University, Melbourne, VIC, Australia; and ETH Zurich, Zurich, Switzerland. He was an assistant professor of image understanding with TU Darmstadt, Darmstadt, Germany, in 2009. Since 2010, he has been a tenured professor of photogrammetry and remote sensing with ETH Zurich. His research interests include remote sensing, photogrammetry, computer vision, and machine learning.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/