Temporal Attention Multi-Resolution Fusion of Satellite Image Time-Series, applied to Landsat-8 and Sentinel-2: all bands, any time, at best spatial resolution

Julien Michel, Jordi Inglada

Highlights

Temporal Attention Multi-Resolution Fusion of Satellite Image Time-Series, applied to Landsat-8 and Sentinel-2: all bands, any time, at best spatial resolution

Julien Michel, Jordi Inglada

- A general formulation of the fusion of Satellite Image Time Series (SITS) is proposed
- A new training strategy and Neural Network architecture solves this general problem
- This solution avoids several unrealistic assumptions found in the litterature
- The pre-trained model is evaluated on 4 usual tasks using Landsat and Sentinel-2 SITS
- This single model is on-par or better than ad-hoc existing models, with more benefits

Temporal Attention Multi-Resolution Fusion of Satellite Image Time-Series, applied to Landsat-8 and Sentinel-2: all bands, any time, at best spatial resolution

Julien Michel[a,*], Jordi Inglada[a]

[a] *CESBIO, Université de Toulouse, CNES, CNRS, INRAE, IRD, UT3, 18 avenue Edouard Belin, BPI 2801, TOULOUSE Cedex 9, 31401, France*

Abstract

This paper introduces a general formulation for the fusion of Satellite Image Time Series (SITS) of variable length from several sensors at different spatial resolutions and acquisition times over the same geographical area. In this formulation, all the spectral bands from all the input sensors are predicted at the best input spatial resolution, and at any observed or non-observed acquisition time requested. To address this general problem, an advanced Masked Auto-Encoder training strategy is proposed, utilizing two new loss functions: a Linear-Regression Learned Perceptual Image Similarity term to favor high spatial frequency details, and a mask-constrastive term to ignore clouds and other non-informative areas in the input data. This strategy is applied to the training of Temporal Attention Multi-Resolution Fusion of Satellite Image Time-Series (TAMRFSITS), a novel Deep Learning architecture designed to implement the proposed general formulation. Experiments with joint Landsat-8/9 and Sentinel-2 time-series were conducted on four different tasks from the literature and demonstrate that a single pre-trained TAMRFSITS is on par or better than existing ad-hoc methods. Moreover, the proposed method relaxes unrealistic assumptions routinely found in the literature, including: same or similar spectral bands in different sensors, same-day acquisitions, and scale-invariance of the relationship between high and low resolution images. To the best of our knowledge, our method is the first to achieve this range of capabilities with a single model, without making any of these

[*] Corresponding author

assumptions. The complete source code for training and experiments is available here: https://github.com/Evoland-Land-Monitoring-Evolution/tamrfsits.

Keywords: Super-Resolution, Spatio-Temporal Fusion, Sharpening, Gap-Filling, Sentinel-2, Landsat, Transformer

1. Introduction

Global coverage of the Earth surface with Satellite Image Time Series (SITS) is a critical component of the continuous monitoring of our planet Essential Climate Variables Bojinski et al. (2014) and Essential Biodiversity Variables Jetz et al. (2019). The Landsat series alone provides 50 years of such constant monitoring Wulder et al. (2022), and for ten years now Sentinel-2 has complemented the Landsat archive with higher revisit, higher spatial resolution images. SITS are used to monitor vegetation Misra et al. (2020), derive Land Cover and Land Use maps Phiri et al. (2020), control the European Common Agricultural Policy Vajsová et al. (2020) or monitor Water Bodies Bhangale et al. (2020).

Spatial and temporal resolutions of SITS are important parameters that define the reachable scale of analysis both in space and time. The revisit time is even more critical for passive optical sensors, where cloud occurrences obliterates a vast proportions of the observations Wilson and Jetz (2016). On the other end, there is an increasing number of orbiting sensors that could complement each other to increase the revisit time and the spatial resolution. In this paper, we consider the case of two or more contemporary sensors with global coverage and regular revisit, complementary revisit cycles and spectral bands, and different spatial resolutions. There are a few combinations of such sensors already flying, the most popular being Landsat 8/9 with Sentinel-2, because of their complementary revisit cycles Li and Chen (2020). Another already available combination is PlanetScope and Sentinel-2 Latte and Lejeune (2020); Sadeh et al. (2021). In a near future, possible combinations will include Trishna Lagouarde et al. (2018) and LSTM Bernard et al. (2023) with Sentinel-2 or Landsat, as well as Sentinel-2 NG or Landsat-Next. There is therefore a growing need for methods that can jointly leverage those continuous sources of earth monitoring. Ideally one would like to transform all observations from all sensors into measurements as seen by a ubiquitous sensor: all wavelengths, any time, at best spatial resolution.

1.1. Existing works

In the literature, this general problem is partly addressed by several families of sub-problems, with different hypotheses and constraints. Those sub-problems and the methods that have been proposed to solve them are detailed in section 1.1.1. Another corpus of work focuses on producing harmonized datasets for identified sets of sensors. To that end, they propose end-to-end processing frameworks including pre-processing and sensor-specific features, in addition to leveraging methods from the above literature. They are summarized in section 1.1.2.

1.1.1. Taxonomy of methods

The relevant methodological subdomains are summarized in Fig. 1. They include Single and Multi-Image Super-Resolution, Band-Sharpening, Spatio-Temporal Fusion and Temporal Modeling, all of which are detailed in the following sections. Those works usually focus on machine learning methodological developments within the constraints of existing datasets or subdomain definitions.

Single and Multi Image Super-Resolution. In Single or Multi-Images Super-Resolution (SISR or MISR) Anwar et al. (2020); Liu et al. (2022), images from a higher resolution sensor are used to learn a model improving the spatial resolution of one or several images acquired by a lower resolution sensor. SISR and MISR are inherently ill-posed, since there can be many High-Resolution (HR) images corresponding to the same observed Low-Resolution (LR) image(s). For this reason, successful SISR and MISR usually rely on adversarial training Wang et al. (2018); Ledig et al. (2017) to capture the target data manifold structure during training. Both SISR Galar et al. (2019); Nguyen et al. (2023); Salgueiro Romero et al. (2020); Pouliot et al. (2018) and MISR Märtens et al. (2019); Okabayashi et al. (2024); Molini et al. (2019); Ibrahim et al. (2025) have been applied to remote-sensing imagery. Models are usually trained by using either simulated or cross-sensor datasets, each having their own strengths and weaknesses, as analysed in a previous work Michel et al. (2025). Though vastly investigated in the

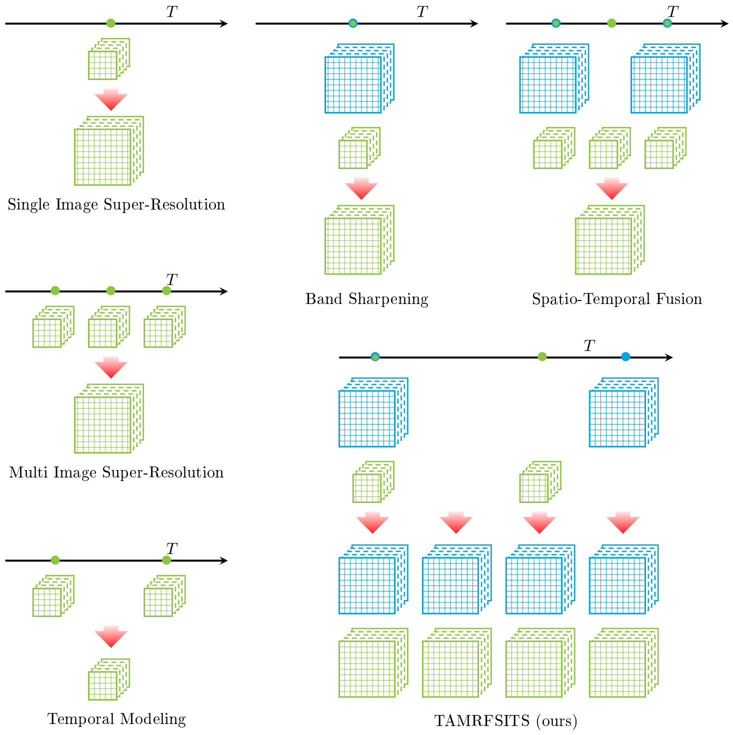


Figure 1: Overview of capabilities of each relevant methodological subdomain, as compared to our proposed method. Input SITS are represented along a time-line showing the acquisition times. The Low Resolution (LR) SITS are represented by small green data-cubes, while the High Resolution (HR) SITS are represented by large blue data-cubes. Single Image Super-Resolution predicts the LR SITS bands at the same acquisition time, but at better spatial resolution. Multi-Image Super-Resolution predicts the LR SITS bands at a single acquisition time but at better spatial resolution, using multiple acquisitions in the input LR SITS. Temporal modelling provides prediction of the LR SITS bands at unobserved acquisition times, without changing the spatial resolution. Band Sharpening provides LR SITS bands at better spatial resolution and at same acquisition times, using simultaneous HR SITS as an auxilliary input. Spatio-Temporal Fusion leverage simulateneous observation of LR SITS and HR SITS in order to predict LR bands at better spatial resolution, at acquisition dates when only LR bands are observed. Finally, our proposed methods process joint HR and LR SITS and predict all HR and LR bands at the HR resolution for any acquisition time.

literature, SISR and MISR are not well suited for leveraging the synergies between two global coverage sensors of different spatial resolutions, since they do not integrate the prior HR information given by HR observations at other dates, which always exists in the case of global coverage missions. They can not overcome missing data in input images because it lacks temporal context. Finally, it requires HR reference data with similar spectral bands and even same-date acquisitions in case of cross-sensor datasets. This explains why joint Super-Resolution and sensor translation is not represented in Fig. 1: if a different HR sensor is often used to train the super-resolution model, authors usually try to avoid learning sensor-to-sensor differences Michel et al. (2025), and consider that the output of the super-resolution model should be the same sensor as the input with a better spatial resolution. An exception to this is the work of Sambandham et al. Sambandham et al. (2024), which is covered in section 1.1.2.

Band-Sharpening. Band-sharpening Firozjaei et al. (2022), also known as pan-sharpening Vivone et al. (2014); Ciotola et al. (2025), or reference-based Super-Resolution Su et al. (2025); Lutio et al. (2019) depending on the scientific community, aims at using an auxiliary near-simultaneous High Resolution image (often from the same satellite) to improve the spatial resolution of an image acquired by a lower resolution sensor. As opposed to SISR, where HR images are used as a target in the model supervision, in spatial sharpening HR images are used as an auxiliary input to the model. While adding HR inputs certainly helps the model to improve the spatial resolution, there is no longer an HR reference to train the model against. To overcome this problem, researchers in various remote sensing fields have been using the scale invariance hypothesis, enabling the use of Wald's protocol Wald et al. (1997), which is well illustrated in Fig. 2 of Palsson et al. (2018). It consists in downscaling by a factor k both input and reference data, training the model to perform $r/k \rightarrow r$ resolution improvement, and then applying the trained model to perform $r \rightarrow kr$ resolution improvement by assuming scale invariance. This practice is commonplace in thermal sharpening, a branch of spatial sharpening addressing the Thermal Infra Red or Land Surface Temperature (LST) images Gao et al. (2012); Granero-Belinchon et al. (2019), but it is also used in other band-sharpening applications when sensor is equipped with different spatial resolutions across bands Palsson et al. (2018); Salgueiro et al. (2021).

Several studies show that the scale invariance hypothesis isg unrealistic and can lead to wrong estimation of local textures Nguyen et al. (2022); Merlin et al. (2010). In thermal sharpening, most methods only learn the mapping from degraded HR images to LR images, and then use the learnt model on full HR resolution. During inference, the model only sees the HR image and can not use local LR observations. This imposes to learn a new model for each pair of HR and LR mapping. In order to make sure that predicted dynamic is faithful to the LR observation, residuals between downsampled prediction and LR observation are often injected to form the final prediction. Nevertheless, the residual signal is inherently a LR signal and may hinder the resolution improvement achieved by the model.

Band-sharpening seems more equipped that SISR and MISR to solve the problem at stake: it supports LR bands without HR reference, and in some cases might produce gap-less outputs, if several HR acquisitions are used to circumvent clouds occurrences Shao et al. (2019). Yet it relies on the unrealistic scale invariance hypothesis, and is essentially mono-temporal: it can not extrapolate unseen dates or learn the temporal dynamics.

Spatio-Temporal Fusion. Spatio-Temporal Fusion (STF) Belgiu and Stein (2019); Xiao et al. (2023) is a set of methods derived from the initial Spatial and Temporal Adaptive Fusion Model (STARFM) Gao et al. (2006). STARFM models the relationship between concomitant Landsat and MODIS surface reflectances at the same location, and extrapolates it to a nearby date where only MODIS is observed. It requires LR and HR simultaneous images both before and after this target date. Model-based improvement have followed Belgiu and Stein (2019); Zhu et al. (2018), most of which extend STARFM or try to overcome its limitations by using more complex models, without changing or extending this paired configuration. More recently, the model-based method has been traded for a parametric Deep-Learning (DL) model, including convolutional layers Tan et al. (2018, 2019); Liu et al. (2019); Tan et al. (2022), adversarial training Zhang et al. (2024); Xie et al. (2024), transformers Wu and Huang (2024); Yang et al. (2022); Li et al. (2022) and vision transformers Chen et al. (2022).

It must be stressed than though the DL models proposed in the literature became more complex, very few of them break free from the limitation of requiring one or two pairs of simultaneous LR and HR acquisitions. Notable exceptions are Goyena et al. (2023) and Zhang et al. (2024) where unpaired STF is investigated. They also do not extend the model temporal capability beyond two dates. Additionally several of these methods actually require the scale invariance hypothesis for training, and some of them also require retraining for each target date. Finally, it should be stressed that the model is usually not informed of the acquisition dates, which prevents the learning of temporal modelling and the generalization to other time intervals that were not seen during training.

Temporal Modeling. Temporal modeling refers to the interpolation or forecasting of SITS Moskolaï et al. (2021). One strong driver of such research is to mitigate the occurrence of clouds in passive optical data in order to generate gapless SITS Li et al. (2021), and most methods handle SITS from a single sensor, with some notable exceptions Roy et al. (2008); Irigireddy and Bandaru (2025). In Liu et al. (2024), a transformer is used in the temporal dimension to reconstruct cloudy pixels. The obvious limitation of temporal modeling is that it does not address the spatial and spectral part of the problem at stake.

1.1.2. Producing harmonized datasets

In addition to those methodological subdomains, a corpus of work aims at leveraging those works to address the broader problem of producing harmonized datasets from SITS of different sensors. This includes data pre- and post-processing as well as production strategies.

Some of them are in operational stage, such as the Harmonized Landsat and Sentinel-2 dataset (HLS) Claverie et al. (2018), where differences between sensors and acquisition conditions are carefully processed into a consistent dataset. Yet HLS does not improve the spatial resolution of Landsat observation, and instead provides matching 30-meter Sentinel-2 observations. A competing operational method aiming at the same goals is Sen2like Saunier et al. (2022) where an additional fusion step using a simple residual compensation yields 10-meter Landsat predictions.

Aside from HLS and Sen2like, other more prospective works include Wang et al. (2017), in which Wang et al. propose to use Area-To-Point Regression Kriging (ATPRK), a standard STF method based on geo-statistics, to produce 10-meter Landsat images of corresponding Sentinel-2 bands at times when only Landsat is observed. In Shao et al. (2019), Shao et al. employ the same idea but replace ATPRK with the Super Resolution Convolutional Neural Network Dong et al. (2015) , relying on the scale invariance hypothesis and Wald's protocol for training and evaluation. This idea is also used in Latte and Lejeune (2020) to fuse Sentinel-2 and PlanetScope imagery into a 2.5m Sentinel-2 product, relying again on the scale invariance hypothesis. Additionally, their methodology requires retraining the model for each Sentinel-2 image because of the varying number of input PlanetScope images. The combination of PlanetScope with Sentinel-2 is also explored in Sadeh et al. (2021), using simple rule-based fusion and linear regression in the context of producing Leaf Area Index high resolution time-series. In Luo et al. (2018) the authors proposed STAIR, later refined into STAIR 2.0 Luo et al. (2020), which is a complete framework for the fusion of SISTS from multiple sensors sharing similar spectral bands. The methodology leverages same-day observations to form a series of differences between HR and spatially up-sampled LR differences. Those differences are then interpolated at acquisition times when only LR is observed, and added to the up-sampled LR observations. STAIR includes many pre-processing steps, including spatial registration, spectral adjustment as well as correction of missing data in Landsat due to faulty detector array. A similar idea to STAIR is explored in Wang et al. (2024) where high resolution details are obtained by means of the Smoothing-Sharpening Image Filter. In Mukherjee and Liu (2023), GAN-based SISR is employed to directly predict High Resolution Sentinel-2 bands from low resolution Landsat bands. In Chen et al. (2021), the same idea is applied to the super-resolution of historical Landsat data, using the overlapping period with Sentinel-2 for training, and in Sambandham et al. (2024), albeit without adversarial training. In Chang et al. (2025), the authors leverage the Enhanced Deep Convolutional Spatio-Temporal Fusion Network Tan et al. (2019) in order to fuse Landsat, MODIS and Sentinel data. Most of those works require a substantial amount of cross-calibration, assuming that spectral bands are similar and that same-day acquisitions exist. Most of them have limited temporal modelling, and are unable to produce gap-less output or extrapolate unseen dates for instance.

1.2. Contributions

As shown in the previous section, those different families of methods only address a sub-part of the following more general problem: predict all bands from all-sensors at best observed spatial resolution and for any acquisition time, given the observed SITS from two or more sensors over the same area, without any further assumptions. We hypothesize that failing to address the big picture not only restricts the capabilities of those methods, but also requires additional unrealistic assumptions and limitations to solve the sub-part of the problem at stake. Capabilities, limitations and unrealistic assumptions of existing methods are summarized in Table 1.

In this paper, we first introduce a mathematical formulation of the general problem in section 2.1. We then propose a deep-learning architecture (section 2.2) and a training procedure (section 2.3) that allow to solve it by means of Self Supervised Learning (SSL). Our proposed model, called Time Attention Multi-Resolution Fusion of Satellite Image Time Series (TAMRFSITS) has full capabilities and none of the limitations and unrealistic assumptions of existing families of methods. Fig. 2 illustrates some of those capabilities on our testing dataset which will be introduced in section 3.1. Though the proposed model is not limited to a specific pair of sensors and can even handle more than two sensors, we demonstrate its properties through the fusion of Sentinel-2 and Landsat-8 time-series in section 3.

2. Proposed method

2.1. Problem Formulation

For the sake of simplicity, the problem formulation is presented for two SITS, one with high spatial resolution R called S_{HR} and one with low spatial resolution r called S_{LR}. In addition, we will assume that r is an integer multiple of R:

$$r = kR, k \in \mathbb{N}, \tag{1}$$

Table 1: Capabilities and limitations of the different categories of methods. Temporal refers to Temporal Modeling (see section 1.1.1), SR refers to Super-Resolution (see section 1.1.1), Sharpening refers to Band Sharpening (see section 3.3.2), and STF refers to Spatio-Temporal Fusion (see section 3.3.3). Parents (✓) indicate that some methods in the literature partially achieve the capability or solve the limitation or unrealistic assumption.

Category of methods	Temporal	SR	Sharpening	STF	Ours
Capability					
Improves spatial resolution		✓	✓	✓	✓
Supports bands without HR reference			✓	✓	✓
Produces gap-less outputs	✓		(✓)	(✓)	✓
Extrapolates unseen dates	✓				✓
Learns temporal dynamic	✓				✓
Limitation solved					
Does not require similar bands in sensors			✓		✓
Does not require same-day acquisitions	✓		(✓)	(✓)	✓
Does not use scale invariance hypothesis	✓	✓			✓
Is not limited to pairs or triplets	✓	✓			✓
Does not require no-data masks as inputs	(✓)				✓
Does not require training for each target date	✓	✓	(✓)	(✓)	✓

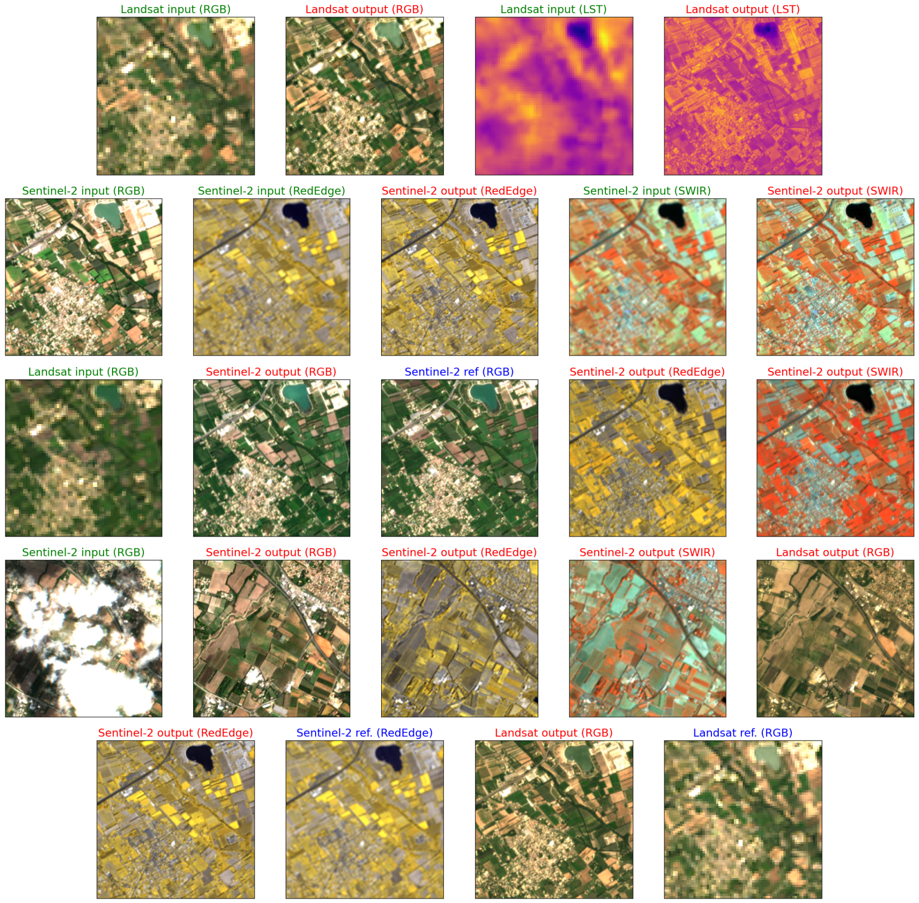


Figure 2: Illustration of the TAMRFSITS model capabilities over Area Of Interest 31TFJ for year 2022 (from the test set, see section 3.1). The model receives 25 Landsat observations and 40 Sentinel-2 observations as inputs (highlighted in green), and is asked to predict Landsat and Sentinel-2 at different query dates (highlighted in red). For some query dates, additional Sentinel-2 or Landsat images (e.g. not part of the model inputs) serve as reference data to compare the prediction with (highlighted in blue). Depending on the availability of clear or cloudy Sentinel2 or Landsat image in the input, several cases are presented. From top to bottom: row 1 shows Landsat prediction for a clear Landsat input date, row 2 shows Sentinel-2 predictions for a clear Sentinel-2 input date, row 3 shows Sentinel-2 prediction compared to true Sentinel-2 image for a clear Landsat input date, row 4 shows Sentinel-2 and Landsat prediction for a cloudy Landsat input date, and row 5 shows Sentinel-2 prediction compared to true Sentinel-2 image for a date when neither Landsat nor Sentinel-2 are seen by the model. Note that the model does not use cloud masks.

where k is the resolution factor between both SITS. The spatial sampling grid of the LR SITS can be defined as $R_{LR} = R(x_0, y_0, h, w, r)$, with:

$$R(x_0, y_0, h, w, r) = \Big\{x_0 + (i+0.5)r, i \in 0, \ldots w-1\Big\} \times \Big\{y_0 + (j+0.5)r, j \in 0, \ldots h-1\Big\}, \quad (2)$$

where x_0 and y_0 denotes the coordinates of the upper-left corner of the grid, and $h \in \mathbb{N}$ and $w \in \mathbb{N}$ denotes the integer height and width of the image. Conversely, the spatial sampling grid of the HR SITS can be written $R_{HR} = R(x_0, y_0, H, W, R)$, with $H = kh$ and $W = kw$. The acquisition times of the LR SITS can be defined as:

$$T_{LR} = \Big\{t_n, n \in 0, \ldots N_{LR} - 1\Big\}, \quad (3)$$

where $N_{LR} \in \mathbb{N}$ is the number of acquisitions in LR SITS. Acquisition times for HR SITS are likewise noted T_{HR}. Finally, $\mathrm{C_{LR}}$ denotes the set of spectral bands in LR SITS, and $\mathrm{C_{HR}}$ denotes the set of spectral bands in HR SITS, without any further assumption on possible matching bands between C_{LR} and C_{HR}. With those notations, both the HR and LR SITS can be seen as a collection of measurements sampled from a universal observation function $F(x, y, r, t, c)$ of spatial location (x, y), spatial resolution r, acquisition time t and spectral band c:

$$S_{LR} = \Big\{F(x, y, r, t, c) \text{ with } (x, y), t, c \in R_{LR} \times T_{LR} \times C_{LR}\Big\}, \quad (4)$$

$$S_{HR} = \Big\{F(x, y, r, t, c) \text{ with } (x, y), t, c \in R_{HR} \times T_{HR} \times C_{HR}\Big\}. \quad (5)$$

For the sake of computation as well as for enforcing spatial and temporal inductive biases, collections of measurements S_{LR} and S_{HR} are usually organized into tensors of shape $[w, h, N_{LR}, \#C_{LR}]$ (resp. $[W, H, N_{HR}, \#C_{HR}]$), where $\#C_{LR}$ (resp. $\#C_{HR}$) denotes the number of elements in C_{LR} (resp. C_{HR}). Let T_{query} denote any other set of acquisition times. In this paper, we aim at building a parametric function $\Phi(S_{LR}, S_{HR}, T_{query} \mid \Theta)$ as well as a training procedure to derive its optimal parameters $\Theta^{\star}$ so that Φ is able to infer measurements for all bands in $C_{all} = C_{HR} \cup C_{LR}$ at every query acquisition time in T_{query} and sampled on high spatial resolution sampling grid defined by R_{HR}:

$$\Phi(S_{LR}, S_{HR}, T_{query} \mid \Theta^{\star}) \simeq \left\{F(x,y,r,t,c) \text{ with } (x,y), t, c \in R_{HR} \times T_{query} \times C_{all}\right\}. \quad (6)$$

Note that the output of Φ can be further separated into two SITS with C_{HR} and C_{LR} channels:

$$\Phi(S_{LR}, S_{HR}, T_{query} \mid \Theta^{\star}) = (S^{\star}_{LR}, S^{\star}_{HR}), \quad (7)$$

with

$$S^{\star}_{LR} \simeq \left\{F(x,y,r,t,c) \text{ with } (x,y), t, c \in R_{HR} \times T_{query} \times C_{LR}\right\}, \quad (8)$$

$$S^{\star}_{HR} \simeq \left\{F(x,y,r,t,c) \text{ with } (x,y), t, c \in R_{HR} \times T_{query} \times C_{HR}\right\}. \quad (9)$$

2.2. Architecture

In this work, function Φ is implemented through a deep-learning architecture called Temporal Attention Multi-Resolution Fusion of Satellite Image Time Series (TAMRFSITS), leveraging residual Convolutional Neural Networks (CNN) in the spatial dimensions and a Transformer in the temporal dimension. The overall architecture follows a classical encoder - decoder scheme, as presented in Fig. 3. Both S_{LR} and S_{HR} are fed into the TAMRFSITS encoder which outputs a latent representation S_{latent} comprising D features, with shape $[W, H, \#T_{LR} + \#T_{HR}, D]$. It is important to note that unlike many works in auto-encoders, the time dimension is not projected to a fixed size, and retains the initial number of observations $\#T_{LR} + \#T_{HR}$. To obtain predictions for T_{query} acquisition times, both T_{query} and latent representation S^{latent} are fed into the TAMRFSITS decoder, which outputs $S^{\star}_{LR}$ and $S^{\star}_{HR}$. In the following sections, the architecture of the encoder (section 2.2.1) and the decoder (section 2.2.2) will be detailed. Intuitively, TAMRFSITS can be seen as the combination of a Single Image Super-Resolution network in the spatial dimension and a Transformer in the temporal dimension.

2.2.1. Encoder

The workflow of the TAMRFSITS encoder is detailed in Fig. 4 and is composed of two sequential stages. First, each acquisition date of S_{LR} is processed independently

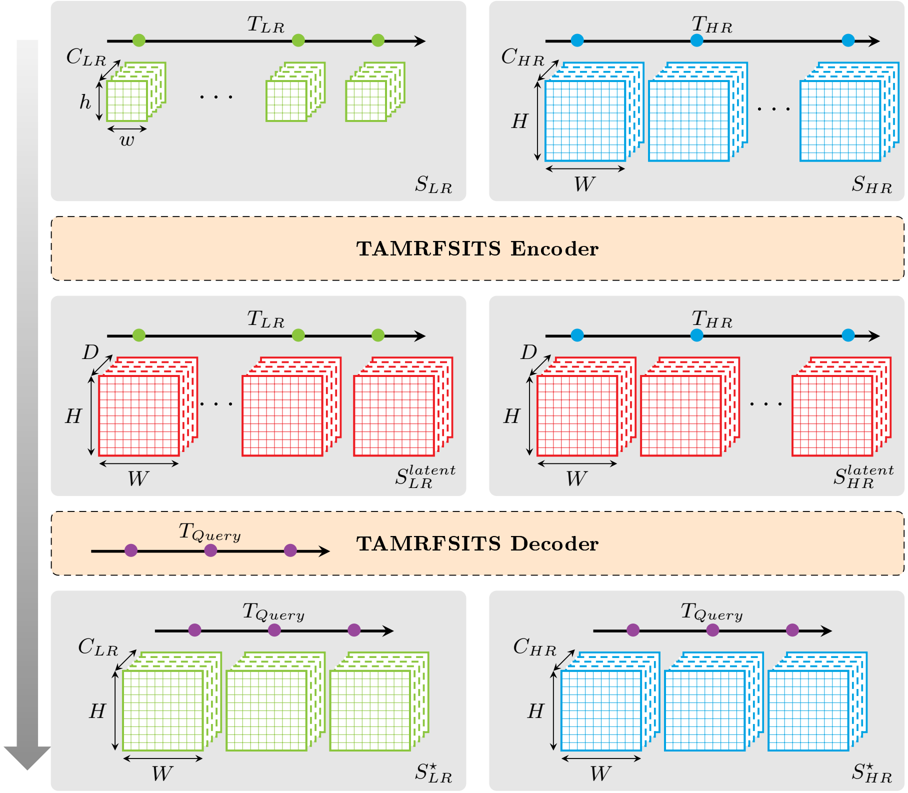


Figure 3: Overview of the proposed method. S_{LR} (in green) and S_{HR} (in blue) are fed into the TAMRFSITS encoder, which outputs latent representation S^{latent} (in red). The latent representation is fed into the TAMRFSITS decoder, along with T_{query}, which outputs $S^{\star}_{LR}$ (in green) and $S^{\star}_{HR}$ (in red). The large downward arrow on the left indicates the processing flow, from top to bottom.

by a dedicated spatial LR encoder. The same applies to the acquisition dates of S_{HR}, which are processed independently by a dedicated spatial HR encoder. The resulting latent SITS are then concatenated along their temporal dimension and processed by a transformer encoder in the same temporal dimension. Thus, the TAMRFSITS encoder follows the spatial then temporal encoding paradigm also found in UBARN Dumeur et al. (2024) or SITS-Former Yuan et al. (2022).

Spatial encoder. The date-wise spatial encoders follow the SRResNet architecture introduced in Ledig et al. (2017) and notably used in ESRGAN Wang et al. (2018). It consists in chaining several Residual-In-Residual-Dense-Block (RRDB), followed by an up-sampling operation achieved by bicubic interpolation. The up-sampling is not applied in the HR encoder. With respect to the original architecture of SRResNet, the convolution layers after the up-sampling operation are removed. These final layers will form the spatial part of the TAMRFSITS decoder, presented in section 2.2.2.

The LR spatial encoder outputs F features on a $W \times H$ spatial support from the $\#C_{LR}$ input channels on a $w \times h$ spatial support. On the other hand, the HR spatial encoder outputs F features on a $W \times H$ spatial support from the $\#C_{HR}$ input channels on a $W \times H$ spatial support. Each encoder is therefore specific to the set of spectral bands (C_{LR} or C_{HR}), and to the spatial resolution it processes. In the outputs of both encoders, all dimensions match except for the temporal dimension.

Temporal and modality encoder. For a pixel corresponding to a spatial location, the F features extracted by the date-wise spatial encoders for each date of S_{HR} and each date of S_{LR} are concatenated into a sequence of measurements. To allow the downstream temporal encoder to reason with sensors and acquisition date, the F features are extended with contextual temporal and sensor information. The temporal context is obtained through the traditional positional encoding used by transformers in Natural Language Processing Vaswani (2017). This positional encoding has been extended to temporal positional encoding in many works on SITS, including Dumeur et al. (2024); Yuan et al. (2022); Guo et al. (2024). It alternates sine and cosine functions with varying frequencies:

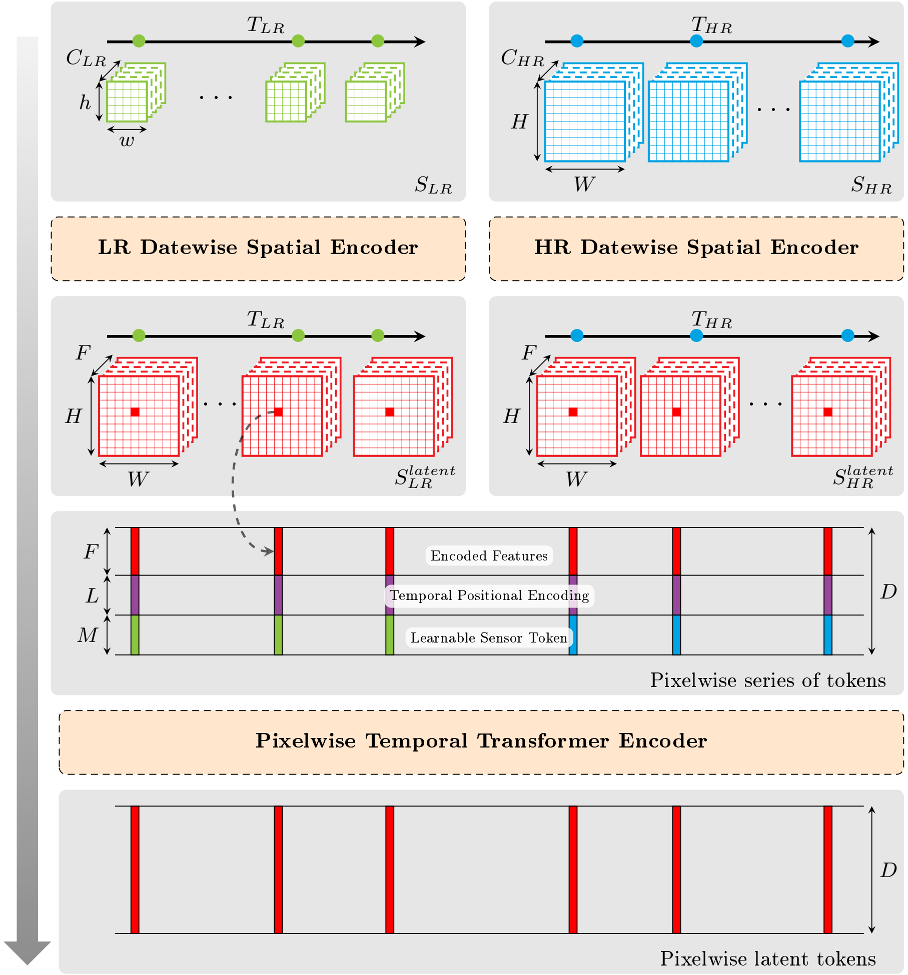


Figure 4: Details of TAMRFSITS encoder. The input SITS S_{LR} (in green) and S_{HR} (in blue) are first passed through separate date-wise spatial encoders based on the Residual-In-Residual-Dense-Block (RRDB) architecture, including spatial up-sampling of S_{LR}. The resulting series S_{LR}^{latent} and S_{HR}^{latent} have matching dimensions, except for the temporal dimension. The resulting F latent features are then extended with temporal positional encoding (L features) and a learnable sensor token (M Features). Therefore, each pixel is converted to a sequence of tokens with a total of D features. This sequence is processed by a pixel-wise transformer encoder along the temporal dimension, forming the latent SITS S^{latent}. The large downward arrow on the left indicates the processing flow, from top to bottom.

$$\Psi(t,l) = \begin{cases} \sin(t/\lambda^{l/L}) \text{ if } l \text{ is even,} \\ \cos(t/\lambda^{bul/L}) \text{ if } l \text{ is odd,} \end{cases} \qquad l \in 0,\dots L-1, \tag{10}$$

where L is the number of temporal positional features and λ is a normalization factor. The L features of the temporal positional encoding are concatenated in the feature dimension with the F features produced by the date-wise spatial encoder. The sensor context is obtained from two dedicated learnable sensor tokens of size M, one for S_{HR} and one for S_{LR}. The corresponding token is also concatenated to the feature dimension. A pixel (i.e. spatial location) is therefore described by a sequence of $\#T_{LR}+$ $\#T_{HR}$ tokens of D features. Note, that since the D feature space includes temporal positional encoding, the order of dates in the latent SITS does not matter.

This pixel-wise sequence of tokens is then processed by a transformer encoder module as described in Vaswani (2017): the sequence passes through several blocks of multi-head self-attention followed by a feed-forward network. The resulting pixel-wise sequences are rearranged into spatial dimensions to form the S^{latent} SITS.

2.2.2. Decoder

The TAMRFSITS decoder operates on the latent SITS S^{latent} produced by the encoder as well as on the acquisition time queries T_{query}. As shown in Fig. 5, it consists of a pixel-wise temporal decoder followed by a date-wise spatial decoder.

Temporal decoder. The temporal decoder mechanism is similar to the decoder used in the cross-reconstruction task of ALISE Dumeur et al. (2024). The first step is to transform the acquisition time queries T_{query} into a sequence of $\#T_{query}$ tokens of D features. A temporal positional encoding of size L is computed using eq. 10, and concatenated to a learnable placeholder of size $F + M$.

The query sequence and the pixel-wise sequences of S_{latent} are then fed to a modified transformer decoder instance: the first self attention block is removed, the S_{latent} sequence serves as keys and values for the multi-head cross-attention block, while the acquisition time query sequence serves as queries. The resulting output pixel-wise sequence can be seen as S_{latent} reconstructed at the acquisition times in T_{query}.

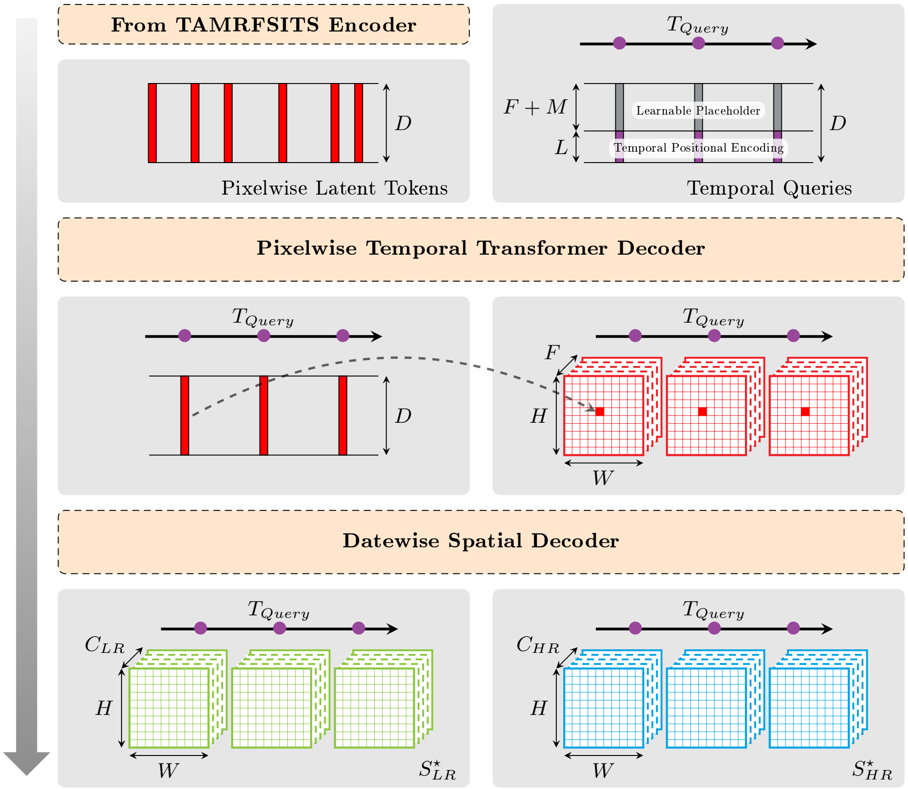


Figure 5: Details of TAMRFSITS decoder. The query acquisition times T_{query} are transformed into a sequence of tokens by means of temporal positional encoding and a learnable placeholder. The query sequence and the latent SITS S^{latent} are processed by a pixel-wise temporal decoder, starting with a cross-attention block, which reconstructs S^{latent} at query acquisition times. The resulting SITS is then passed through a simple spatial decoder in order to predict the resulting $S^{\star}_{LR}$ and $S^{\star}_{HR}$ SITS. The large downward arrow on the left indicates the processing flow, from top to bottom.

Spatial decoder. Following this pixel-wise temporal decoding, the resulting sequences are rearranged into spatial dimensions and inputted to the date-wise spatial decoder. This decoder is fairly simple and corresponds to the final two convolutional blocks of the SRResNet architecture Wang et al. (2018): two convolutional layers, with activation in-between. The spatial decoder outputs $\#C_{LR} + \#C_{HR}$ channels, which can then be split into the two predicted SITS $S^{\star}_{LR}$ and $S^{\star}_{HR}$.

2.3. Training strategy

The problem formulation proposed in section 2.1 can be functionally implemented by the architecture proposed in section 2.2, but finding the optimal parameters $\Theta^{\star}$ requires a training strategy that pushes toward the desired properties of $S^{\star}_{LR}$ and $S^{\star}_{HR}$.

First, predictions should closely match the actual observations that could have been made on that date. This can be achieved through an SSL strategy, and more specifically a Masked Auto-Encoder (MAE) strategy He et al. (2021); Liu et al. (2023); Reed et al. (2022), described in 2.3.1. The reconstruction loss term of the MAE needs to be adapted to the difference in spatial resolution of the input SITS as described in section 2.3.2.

Second, despite the probable presence of clouds in both LR and HR input SITS, no cloud masks are required as input to the model. On the other hand predictions should correspond to ground measurement and be free of clouds or missing data. This is enforced by the use of a novel loss term inspired from contrastive learning, which is presented in section 2.3.3.

Third, all predicted bands should have the best spatial resolution among the input sensors. This is enforced by a dedicated loss term that aims at favoring High Resolution details in all predictions, which is detailed in section 2.3.4.

2.3.1. Masking Strategy

As opposed to Liu et al. (2024) which adopts a purely random masking strategy, the masking strategy for TAMRFSITS is randomly drawn from the pool of strategies described in Table 2. This allows to favor configurations that are unlikely to happen regularly with random masking, and which are of interest for downstream applications.

This includes completely missing the HR or LR SITS, or long term forecasting. For each input SITS, a different strategy is drawn, with random parameters if the strategy has parameters, in order to maximize the input variability.

The network is trained to reconstruct both masked and unmasked dates in S_{HR} and S_{LR}, and thus during training $T_{query} = T_{HR} \cup T_{LR}$. Reconstructing unmasked dates during training is necessary to ensure that the network will behave correctly on those dates during inference. It should be reminded that inferring unmasked dates is of interest with TAMRFSITS, since the model will provide higher resolution versions of unmasked S_{LR} dates and cloud-free predictions of all unmasked inputs. In the following sections, the SITS containing the masked dates will be called S^m_{LR} and S^m_{HR} respectively. The total loss used in order to optimize TAMRFSITS is given by:

$$\begin{aligned} L = &\underbrace{\alpha_m(w^m_{LR}L^m_{LR} + w^m_{HR}L^m_{HR})}_{\text{masked reconstruction term}} + \underbrace{\alpha_c(w^c_{LR}L^c_{LR} + w^c_{HR}L^c_{HR})}_{\text{clear reconstruction term}} \\ &+ \underbrace{\alpha_m(w^m_{LR}\kappa^m_{LR} + w^m_{HR}\kappa^m_{HR})}_{\text{masked contrastive term}} + \underbrace{\alpha_c(w^c_{LR}\kappa^c_{LR} + w^c_{HR}\kappa^c_{HR})}_{\text{clear contrastive term}} + \underbrace{\alpha_s w^c_{HR} L_s}_{\text{spatial term}}, \end{aligned} \tag{11}$$

where superscript ^{c} (resp. ^{m}) denote clear dates (resp. masked dates), L^c_{HR}, L^m_{HR}, L^c_{LR} and L^m_{LR} are reconstruction terms detailed in section 2.3.2, κ^c_{HR}, κ^m_{HR}, κ^c_{LR} and κ^m_{LR} are invalid data contrastive terms detailed in section 2.3.3, L_s is the spatial reconstruction loss term detailed in section 2.3.4, α_m, α_c and α_s are constant weights designed to balance the different loss terms, and w^m_{LR} (resp. w^m_{HR}, w^c_{LR}, w^c_{HR}) is the ratio of masked LR dates with respect to the total number of LR and HR dates in the SITS. The next sections will detail each of those terms.

2.3.2. Reconstruction term

The reconstruction loss terms make use of the Huber loss, introduced in Girshick (2015) and given by:

$$L_1^{smooth}(x) = \begin{cases} 0.5x^2, & \text{if } |x| \le \epsilon \\ \epsilon(|x| - 0.5\epsilon), & \text{otherwise.} \end{cases} \tag{12}$$

The L_1^{smooth} loss behaves like the L_1 loss when differences between predicted and target values are large, limiting the impact of outliers, and like the L_2 when differences

Table 2: Pool of strategies for MAE training.

Strategy	**Description**	**Parameter**
Random	Randomly discard dates from S_{HR} and S_{LR}	Discard rate
Gaps	Generate periods with where no S_{HR} or S_{LR} are available	Length of gaps
No HR	Discard S_{HR} entirely	N/A
No LR	Discard S_{LR} entirely	N/A
Forecast	Discard all dates in S_{HR} and S_{LR} after a given date	Date
Backcast	Discard all dates in S_{HR} and S_{LR} before a given date	Date

are small, which is better for optimization. In the experiments, it has proven to be beneficial over the standard MSE (L_2), especially in early stages of training. Since L_1^{smooth} will operate on standardized values, we choose $\epsilon = 0.1$.

Two loss terms are defined for each of S_{HR} and S_{LR}. Loss terms for S_{HR} are given by:

$$L^c_{HR} = \frac{1}{\sum_{R_{HR}\times T^c_{HR}\times C_{HR}} V^c_{HR}} \sum_{R_{HR}\times T^c_{HR}\times C_{HR}} L_1^{smooth}(S^\star_{HR} - S^c_{HR})V^c_{HR}, \tag{13}$$

$$L^m_{HR} = \frac{1}{\sum_{R_{HR}\times T^m_{HR}\times C_{HR}} V^m_{HR}} \sum_{R_{HR}\times T^m_{HR}\times C_{HR}} L_1^{smooth}(S^\star_{HR} - S^m_{HR})V^m_{HR}, \tag{14}$$

where the sum is performed over all $(x, y), t, c \in R_{HR}\times T^c_{HR}\times C_{HR}$ (resp. $R_{HR}\times T^m_{HR}\times C_{HR}$) and V^c_{HR} (resp. V^m_{HR}) is a binary validity mask that is derived from the metadata of the products and accounts for clouds, shadows, out-of-swath areas, etc. Note that this mask is only used for the loss computation, and therefore only during training, not as input to the network. Using such masks is key to learning quality predictions by avoiding setting cloudy areas as target for the L_1^{smooth} loss.

Because $S^\star_{LR}$ is of spatial resolution R while S^c_{LR} and S^m_{LR} are of coarser spatial resolution r, the formulation of the S_{LR} loss terms is adapted as follows:

$$L^c_{LR} = \frac{1}{\sum_{R_{LR}\times T^c_{LR}\times C_{LR}} V^c_{LR}} \sum_{R_{LR}\times T^c_{LR}\times C_{LR}} L_1^{smooth}((S^{\star}_{LR} * \omega_\sigma) \downarrow_k - S^{c}_{LR})V^c_{LR}, \tag{15}$$

$$L^m_{LR} = \frac{1}{\sum_{R_{LR}\times T^m_{LR}\times C_{LR}} V^m_{LR}} \sum_{R_{LR}\times T^m_{LR}\times C_{LR}} L_1^{smooth}((S^{\star}_{LR} * \omega_\sigma) \downarrow_k - S^{m}_{LR})V^m_{LR}, \tag{16}$$

where ω_σ is a gaussian spatial kernel designed for smoothing the images in $S^{\star}_{LR}$ and $\downarrow_k$ is the nearest-neighbour downsampling operator of factor k. Combining those operators allow to bring high resolution prediction $S^{\star}_{LR}$ back to its original low resolution r.

2.3.3. Invalid data contrastive term

The use of validity masks V^m_{HR}, V^c_{HR}, V^m_{LR} and V^c_{LR} to mask the reconstruction loss terms should theoretically help the temporal transformer to learn that input invalid pixels such as cloud, cloud shadow or missing data are uninformative for the temporal reconstruction and should not be used. However, those masks from Level 2 products have commission and omission errors which might actually prevent to learn to ignore those pixels. To avoid cloud or other invalid pixel leakage into the predictions, we introduce a invalid contrastive term enforcing the prediction of a masked reference pixel to be closer to the closest valid pixel in time than to the invalid reference pixel. This is formulated as a triplet margin loss Balntas et al. (2016):

$$\kappa_{LR} = \max\left\{L_2(S^{\star}_{LR}, S^{valid}_{LR}) - L_2(S^{\star}_{LR}, S^{invalid}_{LR}) + margin, 0\right\}, \tag{17}$$

where L_2 is the MSE, $S^{invalid}_{LR}$ represents invalid pixels of S_{LR}, S^{valid}_{LR} represents the closest valid pixel in time, and *margin* is a user-defined threshold. κ_{HR} is defined likewise.

In practice, a threshold on the time distance to the closest valid pixel is used, so that valid pixels that are too further away in time are not used in the computation of the contrastive loss term.

2.3.4. Spatial term

The L^c_{LR} and L^m_{LR} reconstruction loss terms operate at initial LR resolution. As such, they can not drive $S^{\star}_{LR}$ toward high spatial resolution details, even if $S^{\star}_{LR}$ has the spatial sampling of R_{HR}. The usual solution to this problem is to assume scale invariance and apply Wald's protocol. In our work, this hypothesis is avoided by leveraging the correlation between S_{HR} and S_{LR} SITS in order to favor high spatial frequencies in $S^{\star}_{LR}$.

Indeed, during training $T_{query} = T_{HR} \cup T_{LR}$, and both C_{LR} and C_{HR} bands are predicted for each time in T_{query} (see Eq. 7). Therefore, for each acquisition in T_{HR} (either masked or clear in the MAE strategy), predicted bands in $S^{\star}_{LR}$ can be compared to reference HR images in S_{HR}. However, no assumptions have been made regarding corresponding bands in C_{LR} and C_{HR}. Therefore, S_{HR} can not be directly used to supervise the spatial reconstruction of $S^{\star}_{LR}$.

To overcome this issue, we project the C_{HR} channels from S_{HR} onto the C_{LR} channels of $S^{\star}_{LR}$, by means of a linear regression. Ridge regularization is employed to make the linear regression more robust to degenerated cases. This linear projection is performed separately for each date in T_{HR}, to compensate for the relatively weak estimate provided by linear regression. This is especially true for bands in C_{LR} that do not have a close match in C_{HR} or depend on external factors that are not measured by bands in C_{HR}. The latter case is encountered with the thermal band of Landsat-8, whose primary driver is meteorological. This linear regression is denoted $\Lambda(S^{\star}_{LR}, S_{HR})$.

Despite the date-wise estimation of the linear regression, $\Lambda(S^{\star}_{LR}, S_{HR})$ is still a weak estimate of S_{LR}. In particular, some local trends might not be correctly accounted for, and using $\Lambda(S^{\star}_{LR}, S_{HR})$ for the direct supervision of $S^{\star}_{LR}$ might lead to radiometric distortion of the latter. Since radiometric accuracy is enforced by the reconstruction loss term of Eq. 15, only the high spatial frequency content of $\Lambda(S^{\star}_{LR}, S_{HR})$ should supervise the high spatial frequency content of $S^{\star}_{LR}$. This is achieved through the use of the Learned Perceptual Image Patch Similarity (LPIPS) Zhang et al. (2018). Indeed, in a previous work Michel et al. (2025), we demonstrated that LPIPS is especially well suited for the measurement of difference in spatial frequency content. Furthermore, it has interesting properties such as being relatively blind to radiometric distortion and spatial misalignment. The complete spatial reconstruction loss term is given by:

$$L_{spat} = LPIPS(S^{\star}_{LR}, \Lambda(S^{\star}_{LR}, S_{HR})) \tag{18}$$

3. Experiments

3.1. Dataset

All the experiments have been conducted using a new Landsat and Sentinel-2 dataset called Landsat to Sentinel-2 (LS2S2), which is publicly available[1]. This dataset comprises joint Sentinel-2 and Landsat-8 and Landsat-9 SITS over year 2022. It is composed of 64 Areas Of Interest (AOIs) for the training set and 41 AOIs for the testing set, each covering an area of 9.9x9.9 km² (990x990 pixels for Sentinel-2 and 330x330 pixels for Landsat) with the geographical coverage given by Fig. 6. All dates with more than 25% of clear pixels over the AOI are included in the dataset. This yields a total of 2 984 Sentinel-2 images and 1 609 Landsat-8 and 9 images in the training set. Additional statistics about the number of dates per AOI for each sensor are presented in Table 3. For each sensor, Top-of-Canopy surface reflectance from level 2 products are used. The spectral bands included in the dataset are presented in Table 4. It can be observed that the Landsat sensor does not have Red Edge bands or wide Infra-Red, and conversely Sentinel-2 sensor does not retrieve Land Surface Temperature (LST). In addition to the spectral bands, corresponding quality masks have been used to derive a validity mask for each date of each sensor. This dataset has been gather through the OpenEO API Schramm et al. (2021). Fig. 7 gives an example of SITS extracted from the LS2S2 dataset.

Table 3: Number of 9.9x9.9 km² images in LS2S2 dataset for each sensor and each split. Average, minimium and maximum columns show statistics on the number of acquisition dates per AOI, for each sensor.

	Sentinel2				**Landsat**			
	Total	Average	Min	Max	Total	Average	Min	Max
train	2 984	44	11	94	1 581	26	0	61
test	1 609	39	6	98	736	18	1	56

[1] https://zenodo.org/records/15471890

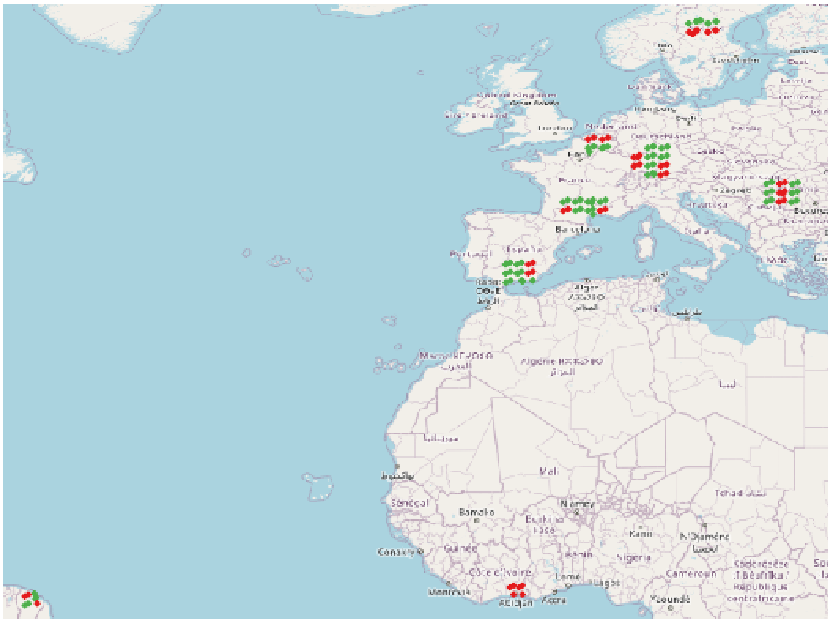
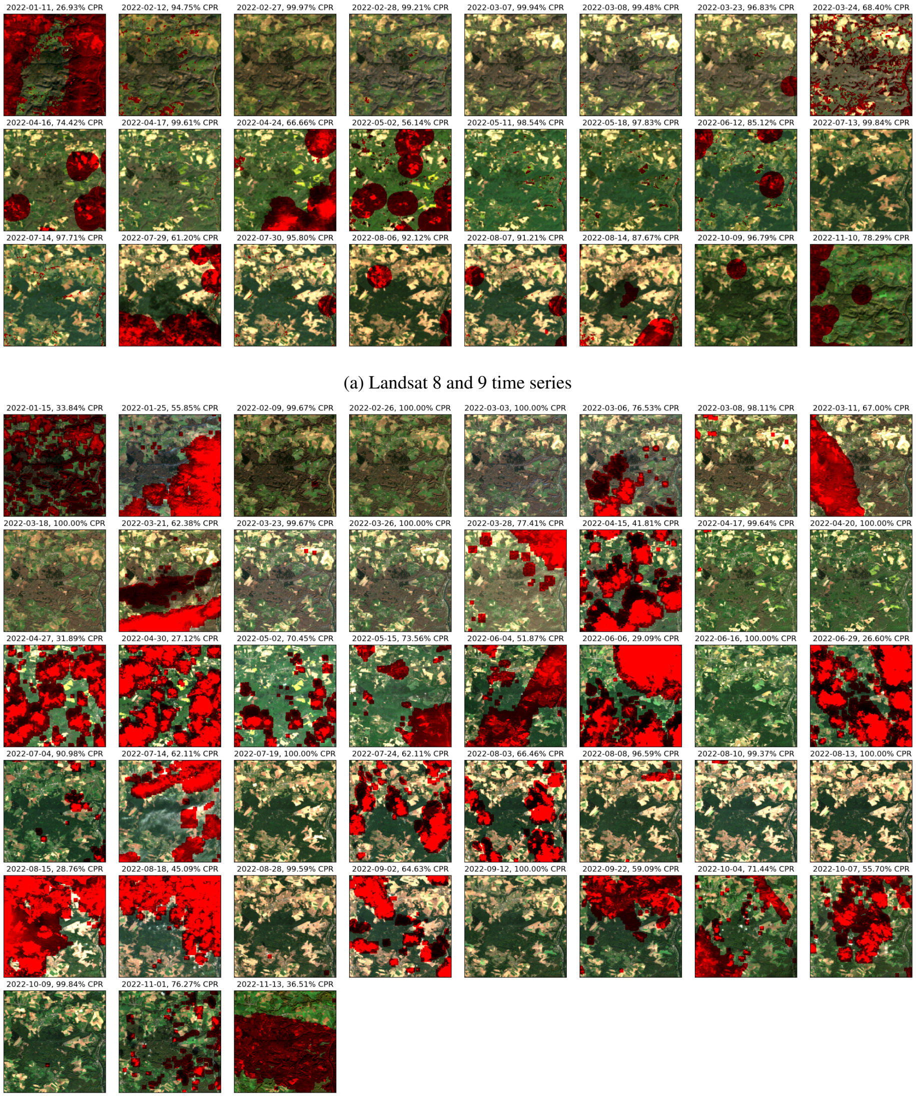


Figure 6: Geographical coverage of the LS2S2 dataset, with training AOIs in green and testing AOIs in red.

Table 4: Spectral bands included in the LS2S2 dataset for each sensor. The table layout allows to highlight which bands are similar between sensors and which bands are only available on one sensor. Despite the variable spatial resolution accross bands for both Landsat and Sentinel-2, for each sensor all bands are up-sampled to the maximum resolution with the bicubic interpolator (10 meter for Sentinel-2 and 30 meter for Landsat).

Sentinel-2		**Landsat**		**Description**
Band	Resolution	Band	Resolution	
		B1	30.	Deep blue
B02	10.	B2	30.	Blue
B03	10.	B3	30.	Green
B04	10.	B4	30.	Red
B05, B06, B07	20.			Red Edge
B08	20.	B5	30.	Near Infra-Red
B8a	10.			Wide Near Infra-Red
B11, B12	20.	B6, B7		Short Wavelength Infra-Red
		B10	100.	Land Surface Temperature



(a) Landsat 8 and 9 time series

(b) Sentinel-2 time series

Figure 7: Example of SITS from LS2S2 dataset training slice, for UTM tile 31UFR. Invalid pixels according to quality masks are highlighted in red.

3.2. Experimental setup

3.2.1. Model and training hyperparameters

The hyperparameters for the TAMRFSITS model have been set experimentally. Table 6 shows the main hyperparameters for each component, as well as their size in number of parameters to optimize. It should be noted that with a total of 2.3M parameters, TAMRFSITS is still a relatively small model compared to the pretrained VGG instance used in LPIPS for instance, which as 16M parameters.

The LS2S2 training split is subdivided into 2052 training patches and 108 validation patches, each of 165x165 pixels for Sentinel-2 SITS and 55x55 pixel for Landsat SITS. Each patch contains the full time-series. TAMRFSITS is trained with the Adam optimizer Kingma and Ba (2014), with one pair of Sentinel-2 and Landsat SITS passed at each step, which can be seen as using a batch size of one. In order to avoid memory errors, if the total number of images for a given step is higher than 50, exceeding images of both sensors are randomly dropped. The masking strategy parameters are summarized in Table 5.

TAMRFISTS is trained for 1000 epochs, for a total of 2×10^6 steps, with validation steps at the end of each epoch. The best model parameters are selected according to the validation loss. The best checkpoint according to the loss measured on the validation patches at the end of each epoch is used as the final model. The initial learning rate is set to 10^{-4}. After a linear warm-up of one epoch, the cosine annealing with warm restarts Loshchilov and Hutter (2016) is used to modulate the learning rate, with an initial period of one epoch and a multiplicative factor of 2. Weights for the different terms of the loss in Eq. 11 were set empirically to α_c=1., α_m=0.5 and $\alpha_\mathrm{s} = 1.$, except for the LST (B10) band of Landsat-8, for which $\alpha_\mathrm{s} = 0.1$. This setting allows to balance the relative importance of the terms. Intuitively, the reconstruction of masked dates is more difficult than the reconstruction of clear dates, hence, masked reconstruction loss terms tend to have higher values. The same applies to the reconstruction of high frequencies of Landsat-8 LST band, which is more challenging than the others due to the lower correlation with the HR bands. All codes use Pytorch Paszke et al. (2019) and run on NDVIA GPU A100 and H100. The total training time is around 10 days on a NVIDIA H100 GPU.

Table 5: Parameters of the MAE strategy described in section 2.3.1. First, a strategy is drawn according to the probability column. Depending on the selected strategy, its parameters are then drawn within the parameter range. U Stands for the uniform distribution. Gaps length and tipping dates are expressed in Day of Year (DoY, 1 for 1st of January and 365 for 31th of December).

Strategy	Probability	Parameter range
Random	0.5	Discard rate ~ U(0.2,0.7)
Gaps	0.25	Length of gaps ~ U(30,90)
No HR	0.0625	N/A
No LR	0.0625	N/A
Forecast	0.0625	Tipping date ~ U(65,300)
Backcast	0.0625	Tipping date ~ U(65,300)

3.2.2. Metrics

This section describes the metrics used to assess the performances of the trained TAMRFSITS model, and compare it to other models from the litterature. Following our previous findings Michel et al. (2025), we retain three qualified metrics, each with a distinctive purpose:

- RMSE measures the faithfulness to the radiometry of the target image at initial resolution,
- BRISQUE is a No-Reference Image Quality metrics that grades the global Image Quality,
- Frequency Restoration (FR) compares the spatial frequency content of the prediction and the target image at initial resolution.

All metrics are measured separately for each band and for each date, either clear (used as model input) or masked (not used as model input).

RMSE. RMSE is computed with respect to the initial resolution of the target data. Since $S^{\star}_{HR}$ has the same spatial resolution R as S_{HR}, $RMSE_{HR}$ has a simple formulation:

Table 6: Summary of model hyperparameters for each component of the TAMRFSITS model. The last column shows the size in number of parameters to optimize for each component.

Component	**Key parameters**		**#Parameters**
LR Spatial Encoder	in channels (C_{LR})	8	797 K
	out channels (F)	64	
	blocks	1	
	up. factor (k)	3	
HR Spatial Encoder	in channels (C_{HR})	10	762 K
	out channels (F)	64	
	blocks	1	
	up. factor	1	
Temporal Encoder	positional encoding (L)	64	435 K
	sensor token (M)	8	
	token size ($D = F + M + L$)	136	
	feed-forward	256	
	nb. layers	3	
	nb. heads	4	
Temporal Decoder	token size ($D = F + M + L$)	136	145 K
	nb. layers	1	
	nb. heads	4	
Spatial Decoder	in channels ($D = F + M + L$)	136	188 K
	out channels ($C_{LR} + C_{HR}$)	18	
Total			2.3 M

$$RMSE_{HR} = L_2(S^{\star}_{HR} - S_{HR}). \tag{19}$$

On the other hand, $S^{\star}_{LR}$ has a spatial resolution of R while S_{LR} has a spatial resolution of $r = kR$. To be able to compute $RMSE_{LR}$, $S^{\star}_{LR}$ is first smoothed by a gaussian kernel and down-sampled to a spatial resolution of R, similarly to loss term presented in eq. 13:

$$RMSE_{LR} = L_2((S^{\star}_{LR} * \omega_{\sigma}) \downarrow_k - S_{LR}). \tag{20}$$

Since $RMSE_{LR}$ is computed at lower resolution r it does not measure the spatial resolution improvement of $S^{\star}_{LR}$. The remaining two metrics are designed for this purpose.

BRISQUE Score. BRISQUE Mittal et al. (2012) is a No Reference Image Quality (IQ) which builds local features mapped to IQ score by a Support Vector Regressor. It is trained on human annotated scores, yielding a score between 0 and 100, 0 being the best IQ. According to Michel et al. (2025), BRISQUE provides a solid criterion to measure the sharpness of the image as well as its overall quality, even if it tends to slightly favor noise. Using a No Reference IQ metric allows providing an insight on the performances that is not biased by the quality of the reference data, such as unmasked clouds, which are present in the LS2S2 dataset.

Frequency Restoration. Frequency Domain Analysis is proposed in Michel et al. (2025) to measure the improvement of the spatial frequency content of super-resolved images. It consists in analysing the Fourier domain Frequency Attenuation Profile ($\mathcal{F}_{\mathcal{AP}}$) for bandwidth $[f_m, f_M]$. Let

$$U_{f_m,f_M} = \{(u,v) : f_m \leq \sqrt{u^2+v^2} < f_M\} \tag{21}$$

denote the set of discrete spatial frequencies (u,v) that lies within a ring defined by f_m and f_M in Fourier plane, $\mathcal{F}_{\mathcal{AP}}$ is given by:

$$\mathcal{F}_{\mathcal{AP}}[P](f_m, f_M) = \frac{1}{\#U_{f_m,f_M}} \sum_{(u,v)\in U_{f_m,f_M}} |\mathcal{F}[P](u,v)|, \tag{22}$$

where $\#U_{f_m,f_M}$ is the number of elements in U_{f_m,f_M}, and $\mathcal{F}[P](u,v)$ is the Discrete Fourier Transform (DFT) of image P. $\mathcal{F}_{\mathcal{AP}}[P](f_m, f_M)$ is successively computed over a set of non-overlapping bandwidth intervals as given by the DFT quantization. The resulting set of values is denoted $\mathcal{F}_{\mathcal{AP}}[P](f_n)$, $n \in [0, N]$ with f_n the central frequency of each frequency intervals. $\mathcal{F}_{\mathcal{AP}}[P](f_n)$ is averaged across batches and dataset. Finally, the normalized logarithmic $\mathcal{F}_{\mathcal{AP}}$ is computed, which gives spatial frequency attenuation in decibels:

$$\mathcal{F}_{\mathcal{AP}}{}^{log}[P](f_n) = 10 \cdot \Big(\log_{10}\left(\mathcal{F}_{\mathcal{AP}}[P](f_n)\right) - \log_{10}\left(\mathcal{F}_{\mathcal{AP}}[P](f_0)\right)\Big). \tag{23}$$

Since there are no High Resolution reference images for S_{LR} and conversely no Low Resolution input images for S_{HR}, we adopt the No-Reference Actual Frequency Restoration metric (Michel et al. (2025), eq. 24), which will be abbreviated Frequency Restoration (FR) in the remaining of the paper:

$$FR_{LR} = \sum \mathcal{F}_{\mathcal{AP}}{}^{log}(S^{\star}_{LR}) - \mathcal{F}_{\mathcal{AP}}{}^{log}((S_{LR}) \uparrow_k)), \tag{24}$$

$$FR_{HR} = \sum \mathcal{F}_{\mathcal{AP}}{}^{log}(S^{\star}_{HR}) - \mathcal{F}_{\mathcal{AP}}{}^{log}(S_{HR}), \tag{25}$$

where $(S_{LR}) \uparrow_k$ is the bicubic up-sampling of factor k of S_{LR}. FR is also expressed in decibels.

3.2.3. Dataset particularities

There are three particularities of the LS2S2 dataset which need to be addressed in the model. First, all Landsat bands are sampled at 30 meter resolution, and all Sentinel-2 bands are sampled at 10 meter resolution. Therefore, the up-sampling factor $k = 3$. This requires a small adaptation of the up-sampling operation in RRDBNet, which usually performs progressive dyadic up-sampling. Second, despite this constant spatial sampling resolution for each sensor, some Sentinel-2 spectral bands have a native resolution of 20 meters, and the LST band of Landsat has a native resolution of 100 meters, as stated in Table 4. This is accounted for in the reconstruction loss term (section 2.3.2) and in the computation of the RMSE metric (section 3.2.2) by using larger gaussian kernels for blurring those bands, independently of the downsampling factor. Finally, the main driver for the Land Surface Temperature band is meteorological. Colder conditions or stronger wind will yield lower LST, while warmer conditions and weaker wind will yield higher LST. Since the TAMRFSITS is not informed by meteorological conditions, trying to predict LST at unobserved Landsat date is meaningless. For this reason, the reconstruction loss term (section 2.3.2) and RMSE metric (section 3.2.2) are only computed for clear Landsat dates for the LST band.

3.3. Results

To our best knowledge, there are no other methods in the literature that provide the same capabilities as TAMRFSITS (see Table 1). In order to provide an overview of the performances of the proposed method with respect to existing ones, several tasks have been identified and assessed on the LS2S2 testing set. They include reconstruction of Landsat and Sentinel2 dates in gaps of one month in section 3.3.1, band-sharpening of Sentinel-2 images in section 3.3.2, Spatio-Temporal Fusion in section 3.3.3 and Thermal Sharpening in section 3.3.4. It is important to note that for all those tasks, a single pre-trained instance of the TAMRFSITS has been used, without any retraining or fine-tuning for the specific task.

For each task, baseline methods from the literature have been selected for comparison. Competing methods have been selected from existing works that address Sentinel-2 and Landsat 8 fusion or spatial resolution enhancement. Only methods for which the implementation was straightforward or for which the authors provided source code and pre-trained weights have been considered.

3.3.1. Gap-Filling

The first task consists in generating one-month gaps in both Landsat and Sentinel-2 SITS for all AOIs in the testing set, following the scheme of Fig. 8. All acquisitions in generated gaps are removed from the model input and kept aside for validation. The model is asked to reconstruct both the missing dates within gaps and the clear input dates, with all bands at 10 meter spatial resolution. Unfortunately, we did not find any DL method with open source code and model weights ready to be used to perform this task on Sentinel-2 and Landsat in the literature. However, a naive algorithm has been implemented for comparison, corresponding to what is usually used as pre-processing for downstream applications requiring temporally and spatially aligned data cubes. This naive method consists in linearly interpolating SITS from each sensor along the temporal dimension, taking into account the cloud and missing data mask. Then, all low resolution bands are spatially interpolated to 10 meter resolution with bicubic up-sampling.

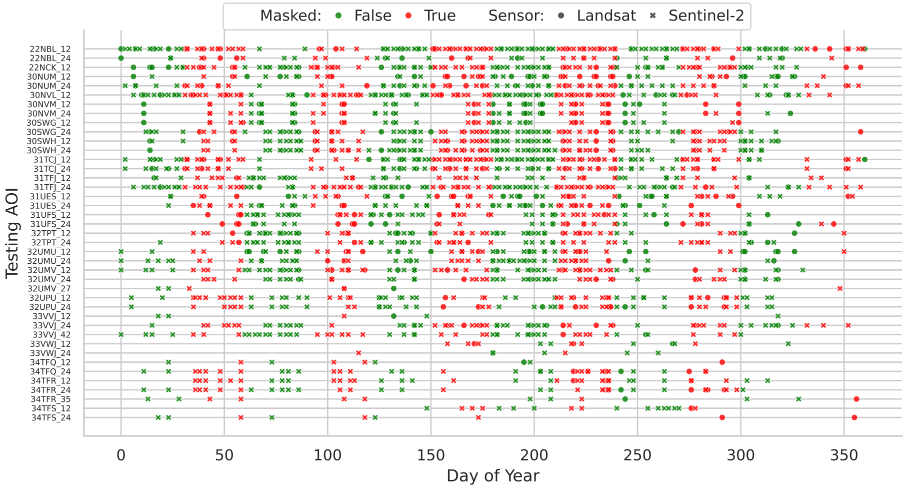


Figure 8: Gaps generated for the gap-filling task evaluation. When masked is True, the acquisition is removed from the model input SITS and used for performances evalution.

Table 7 shows the mean and standard-deviation of the three evaluation metrics, for a subset of Sentinel-2 and Landsat bands (full results are available in appendix 5), separately for masked dates, which are not in the model inputs, and clear dates, which are in the model inputs. Since the RMSE is computed at initial resolution, it is expected that the naive method gives RMSE values that are close to zero for clear dates. When compared to the naive method, TAMRFSITS consistently provides better RMSE on masked dates, with a narrower standard deviation, while being on par with the naive model on clear dates, except for Landsat LST, where RMSE is 4 times higher. However, it provides better Image Quality for all bands, either clear or masked, lowering the BRISQUE score by a vast margin. Finally, spectral bands with an initial resolution higher than 10 meter have positive FR for TAMRFSITS, while the naive method provides FR close to zero. This highlight that our model effectively improves the high spatial frequency content for all bands. This explains why it yields RMSE on clear dates that are slightly worse than the naive method: those dates have more high resolution details. Nevertheless, RMSE values are close to the accuracy of level 2A surface reflectances produced by the MACCS-ATCOR Joint Algorithm (MAJA) Lonjou et al. (2016) processor, and the LST RMSE is also very low with respect to the expected accuracy of Temperature - Emissivity separation algorithm Li et al. (2021).

		RMSE↓				**BRISQUE↓**				**FR↑**			
		Clear		Masked		Clear		Masked		Clear		Masked	
Band	**Method**	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
LS B1	naive	**0.002**	0.001	0.010	0.008	84.6	6.1	84.8	6.0	0.9	1.9	0.7	2.0
	tamrfsits	0.005	0.018	**0.008**	0.004	**42.8**	14.7	**46.1**	19.8	**6.9**	1.8	**7.2**	2.1
LS B4	naive	**0.005**	0.002	0.020	0.018	72.6	3.7	72.6	3.4	0.9	1.9	0.9	2.3
	tamrfsits	0.007	0.018	**0.014**	0.010	**26.6**	10.9	**21.2**	9.0	**7.4**	1.6	**7.6**	1.9
LS B5	naive	**0.006**	0.001	0.031	0.025	71.1	2.5	71.6	2.1	1.4	2.8	1.0	2.8
	tamrfsits	0.012	0.012	**0.027**	0.014	**27.5**	12.2	**13.6**	7.2	**7.0**	1.9	**7.7**	2.0
LS B7	naive	**0.006**	0.001	0.030	0.027	69.9	2.0	70.0	1.8	1.0	2.0	0.9	2.2
	tamrfsits	0.009	0.013	0.020	0.014	**25.6**	12.8	**14.7**	6.0	**7.3**	1.5	**7.7**	1.7
LS LST	naive	**0.460**	0.449			82.3	9.8			1.9	4.5		
	tamrfsits	1.270	0.927			**19.3**	20.0			**3.2**	2.1		
S2 B4	naive	**0.000**	0.000	0.036	0.055	17.5	6.8	18.5	6.7	1.4	2.8	1.2	3.1
	tamrfsits	0.011	0.018	**0.032**	0.049	**17.5**	8.0	**20.9**	7.8	1.4	2.8	1.2	3.3
S2 B6	naive	**0.008**	0.004	0.045	0.046	47.2	4.5	47.9	4.9	1.4	2.2	1.3	2.4
	tamrfsits	0.014	0.017	**0.040**	0.042	**12.5**	6.2	**14.5**	6.2	**5.8**	1.9	**5.7**	2.4
S2 B8	naive	**0.000**	0.000	0.052	0.043	12.9	6.3	13.9	6.0	0.8	2.0	0.7	2.4
	tamrfsits	0.015	0.017	**0.046**	0.039	**11.8**	6.3	**16.0**	6.9	0.8	1.9	0.6	2.4
S2 B8a	naive	**0.009**	0.004	0.049	0.039	47.3	4.9	47.9	5.1	1.3	2.0	1.2	2.2
	tamrfsits	0.016	0.016	**0.043**	0.036	**13.6**	6.6	**14.5**	6.7	**5.5**	1.6	**5.6**	2.1
S2 B12	naive	**0.005**	0.002	0.029	0.019	52.9	6.2	53.8	6.4	1.5	2.2	1.4	2.4
	tamrfsits	0.010	0.007	**0.026**	0.017	**15.8**	10.6	**15.4**	6.3	**6.3**	2.1	**6.0**	2.3

Table 7: Comparison between TAMRFSITS model and naive interpolation on the gap-filling task, where regular gaps of 30 days are masked from the input SITS and kept appart for validation. Only a selection of spectral bands is presented. Full results are available in appendix 5. ↓ (resp. ↑) indicates that the metric should be minimized (resp. maximized). Best mean values for each metric and each band are highlighted in **bold**. **Clear** designate dates for which the bands were observed by the model, and **masked** designates the dates that were removed from the input SITS as described in Fig. 8.

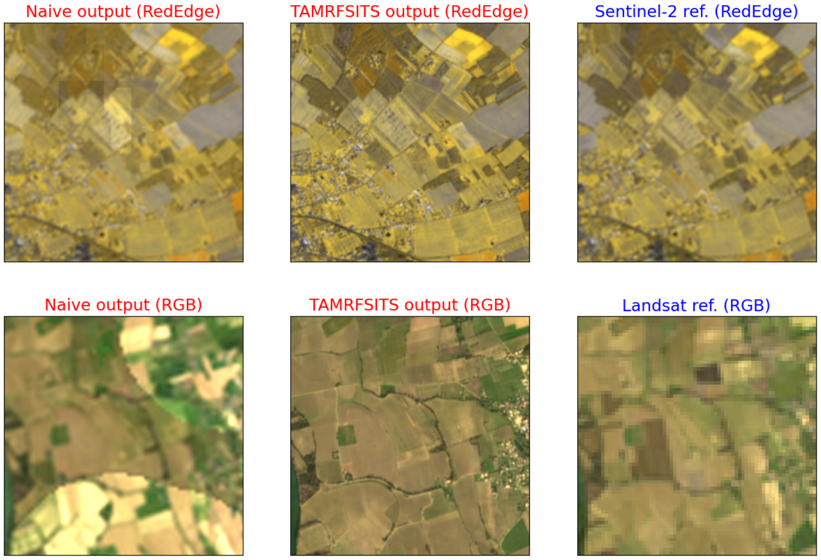


Figure 9: TAMRFSITS compared to naive baseline in interpolating a target date in generated gaps, for tile 31TCJ, for Sentinel-2, 2022-04-10 (top row) and Landsat, 2022-02-10 (bottom row). Sentinel-2 color composition is (B7, B6, B5), while Landsat color composition is natural colors (RGB). Interpolation artifacts caused by mask transition in the naive method are especially visible on the RGB output.

Fig. 9 shows predictions of both TAMRFSITS for a target date located in a gap period for tile 31TCJ. It highlights an important benefit of TAMRFSITS that is not reflected in RMSE: temporal interpolation relies on L2A clouds masks. Depending on those masks, neighboring pixels be interpolated with different dates, yielding artifacts following the mask delineation, which are clearly visible on the bottom row. TAMRFSITS, on the other hand, does not use the cloud masks as input, and therefore yields images that are free from those artifacts. It can also be observed that in addition to predicting consistent surface reflectance values with respect to the reference image, TAMRFSITS also improves the spatial resolution of the spectral bands for both Sentinel-2 and Landsat. In contrast, the naive model only performs spatial bicubic up-sampling, yielding blurry 10 meter resolution data. This highlights the benefits of using TAMRFSITS as both a temporal interpolation method and a spatial resolution enhancement method.

3.3.2. Band-Sharpening

The second evaluation task consist in sharpening the 20 meter Sentinel-2 bands (see Table 4) to 10 meter resolution. TAMRFSITS systematically outputs sharpened predictions for 20 meter Sentinel-2 bands, regardless of whether the query date has been observed by the model or not. In this experiment, it is compared to the DSen2 model Lanaras et al. (2018) , a network that performs similar Sentinel-2 20 meter bands sharpening for a given Sentinel-2 image. DSen2 is applied to each Sentinel-2 acquisition throughout the testing set. On the other hand TAMRFSITS is fed with the full Sentinel-2 SITS for a given AOI, without the Landsat corresponding SITS, and is asked to predict all input Sentinel-2 dates. In addition to sharpening the 20-meter bands, it will therefore also remove clouds and no-data areas. The three evaluation metrics are computed with respect to the input 20 meter spectral bands for both methods, taking into account their validity mask to compute RMSE.

Table 8 shows the comparative performances obtained for a subset of the spectral bands (full results are available in appendix 5), namely one Red Edge band, the Narrow Near Infra-Red band and one of the SWIR bands. It shows that TAMRFSITS is globally on par with DSen2. It tends to yield slightly higher RMSE with respect to the input image, but better Image Quality according to the BRISQUE score, and slightly better sharpening according to the FR score. This is confirmed by the visual inspection of Fig. 10: Dsen2 predictions look less sharp that those from TAMRFSITS, even if both manage to improve on the bicubic up-sampling of the input Sentinel-2 image. The slightly higher RMSE of the TAMRFSITS prediction is barely noticeable. An additional benefit of using TAMRFSITS over DSen2 is of course its ability to seamlessly interpolate cloudy pixels, as demonstrated in Fig. 11.

		RMSE↓		**BRISQUE↓**		**FR↑**	
Band	**Method**	mean	std	mean	std	mean	std
B6	dsen2	**0.003**	0.001	25.34	5.27	4.5	0.5
	tamrfsits	0.013	0.014	**12.86**	6.18	**5.9**	1.9
B8a	dsen2	**0.004**	0.001	24.90	5.47	4.5	0.4
	tamrfsits	0.015	0.014	**13.94**	6.50	**5.6**	1.7
B12	dsen2	**0.002**	0.001	32.18	4.99	4.0	0.7
	tamrfsits	0.010	0.009	**15.78**	9.71	**6.4**	2.0

Table 8: Comparison between TAMRFSITS and DSen2 on the sharpening of Sentinel-2 20m bands, for a selection of bands. ↓ (resp. ↑) indicates that the metric should be minimized (resp. maximized). Best mean values for each band and each metric are highlighted in **bold**. Full results are available in appendix 5.

3.3.3. Spatio-Temporal Fusion

The next task aims at predicting Sentinel-2 at times when only Landsat has been observed, by optionally leveraging Sentinel-2 acquisitions from other time steps. Using the LS2S2 test set, this use case is simulated by removing all Sentinel-2 acquisitions occurring on same date as Landsat from the model input and keeping them as references for validation. The performances are compared to STAIR and Sen2Like, which are rule-based fusion methods, and to Deep-Harmonization Sambandham et al. (2024) and DSTFN Wu et al. (2022), which are DL based methods. All those methods share the same assumption of matching bands between Landsat and Sentinel-2. As such, they do not predict the Red Edge bands, nor the Wide Near Infra Red band (see Table 4).

Sen2like Saunier et al. (2022) is a processing chain providing harmonized surface

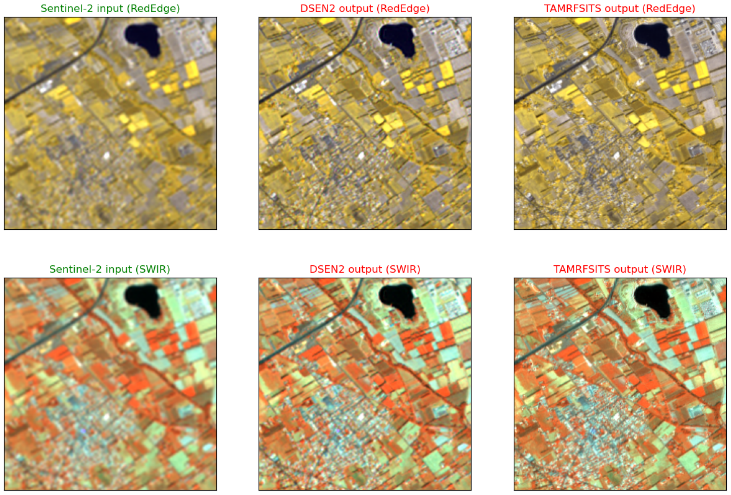


Figure 10: Examples of predictions from DSEN2 and TAMRFSITS over tile 31TFJ, 2022-08-07, for Red-Edge color composition (B7,B6,B5) and SWIR color composition (B8a, B11, B12). Inputs are captioned in green, predictions are captioned in red.

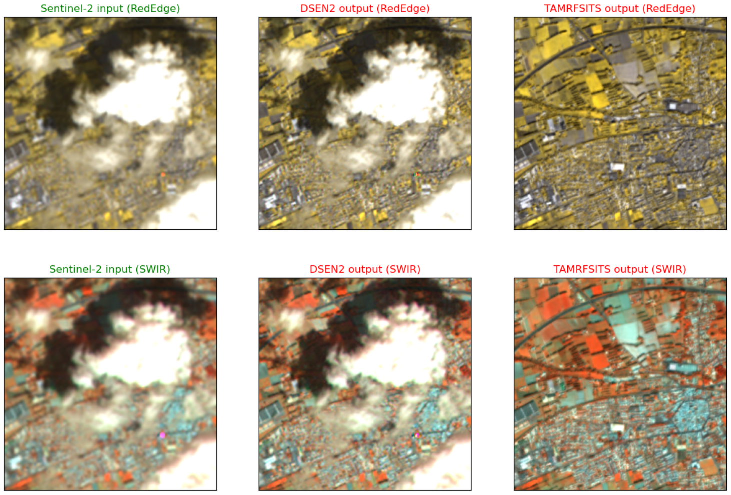


Figure 11: Examples of predictions from DSEN2 and TAMRFSITS over tile 31TFJ, 2022-09-29, where the input Sentinel-2 image is cloudy, for Red-Edge color composition (B7,B6,B5) and SWIR color composition (B8a, B11, B12). Inputs are captioned in green, predictions are captioned in red.

reflectances by means of spatial registration and careful radiative transfer modelling. Though the standard output is produced at 30 meter resolution, an optional fusion rule is proposed in order to up-sample the prediction to 10 meter resolution. For a given Landsat target date, it consists in finding the two closest Sentinel-2 anterior dates with low cloud cover, and build a linear model to interpolate them at the target date. The predicted image is then High Pass filtered to extract the high resolution details, which are added to the target Landsat image. Note that though in Sen2like the 2 dates are searched among the previous predictions in an auto-regressive fashion, in our work they are searched among the available Sentinel-2 dates. As such, we do not compare with the full Sen2like process but only with the temporal extrapolation and spatial fusion methods implemented in Sen2like. STAIR Luo et al. (2018) is another processing chain aiming at merging Sentinel-2 and Landsat images. It first performs a linear gap-filling of the Sentinel-2 SITS and Landsat SITS independently, as well as bicubic up-sampling of Landsat SITS to 10 meter spatial resolution. It then uses same-day pairs in order to build a time-series of differences between Landsat and Sentinel-2 bands. This time-series is linearly interpolated at prediction dates where only Landsat is available, and added to the Landsat image. Deep-Harmonization Sambandham et al. (2024) is an ensemble of five Single Image Super-Resolution Unets trained to map individual Landsat acquisition, including the panchromatic channel, to Sentinel-2 resolution and spectral bands. The ensemble includes instances of the same Unet architecture with two different depths and two different up-sampling layers. Being a SISR approach, Deep-Harmonization does not use Sentinel-2 as input during inference and process the Landsat image independently for each date. DSTFN Wu et al. (2022) is a Spatio-Temporal Fusion model that does not require 2 pairs of same-day acquisitions framing the target date. Instead, DSTFN only requires the Landsat image for target date and an auxiliary Sentinel-2 image at a close date. To compete with these 4 methods, the TAM-RFSITS model can be used in different ways. First, it can process the Landsat images only, in a similar setup as Deep-Harmonization. Second, it can work on Sentinel-2 images only, resorting to the learnt temporal interpolation in order to perform the task. Finally, it can receive all possible Landsat images and all non-simultaneous Sentinel-2 images.

Table 9 shows the performances of the 4 competing algorithms as well as the 3 TAMRFSITS runs with the different input configurations: only LR SITS, only HR SITS and both LR and HR SITS . First, it can be observed that TAMRFSITS systematically yields better RMSE values than the other algorithms. The worst method in terms of RMSE is Sen2like. The Sen2like fusion algorithm expects carefully harmonized surface reflectances between Landsat and Sentinel-2. In this experiment, the fusion process of Sen2like is instead applied to raw L2A surface reflectances, which can explain its poor performance. Following Sen2like closely are the both DL based methods, which also have poor RMSE performances. It can be observed that TAMRFSITS provides good estimate of Sentinel-2 surface reflectance even in the case where only Landsat images are seen by the model (TAMRFSITS no hr). With respect to using only the HR images for the prediction of TAMRFSITS (TAMRFSITS no lr), adding the Landsat acquisition of the target date to the inputs allows to systematically decrease both the mean RMSE and its standard-deviation, which shows that TAMRFSITS makes use of the Landsat low resolution information in the prediction of Sentinel-2 bands. In terms of general image IQ, TAMRFSITS also provides almost always the lowest BRISQUE score, surpassing all methods by a margin of more than 10 BRISQUE points, except for B4 where STAIR is on par with TAMRFSITS. Among the three variants of TAMRFSITS inputs, using only Landsat image yields worse IQ, on par with the other methods, and using the Landsat image of the target date in addition to the Sentinel-2 SITS, yields slightly worse IQ than using only the Sentinel-2 SITS, albeit with lower RMSE. Finally, in terms of spatial resolution enhancement, it can be observed that for the 10 meter bands, all methods manage to retrieve the target resolution (|FR|≈ 0.), to the exception of Deep-Harmonization, which fails to restore the spatial frequency content of a 10 meters image despite being a SISR method at its heart. For 20 meters bands however, only TAMRFSITS and to a lower extent DSTFN manage to provide a spatial frequency content improvement.

Fig. 12 shows sample predictions for all algorithms and provides a visual confirmation of those conclusions. The natural color compositions appear sharp except for Deep-Harmonization, which is both blurry and radiometrically inaccurate. As measured in Table 9, IQ is poor for Sen2like and DSTFN. Deep-Harmonization IQ is on par

		RMSE↓		**BRISQUE↓**		**FR↑**	
Band	**Method**	mean	std	mean	std	mean	std
B4	sen2like	0.031	0.033	29.5	17.2	-0.5	2.8
	stair	0.018	0.010	**18.9**	6.4	0.5	2.1
	dh	0.027	0.015	49.0	6.3	-4.5	2.0
	dstfn	0.027	0.015	27.4	6.2	0.6	1.9
	tamrfsits (no hr)	0.020	0.009	45.9	5.3	-2.7	2.1
	tamrfsits (no lr)	0.017	0.010	19.4	7.3	0.4*	2.1
	tamrfsits (full)	**0.015**	0.008	20.1	7.5	**0.3***	2.1
B8a	sen2like	0.048	0.033	58.4	10.6	-0.3	2.3
	stair	0.033	0.014	48.2	5.2	0.7	1.8
	dh	0.041	0.021	65.3	3.8	-1.9	1.5
	dstfn	0.047	0.033	23.3	5.7	3.3	2.1
	tamrfsits (no hr)	0.031	0.016	42.7	5.0	2.5	1.5
	tamrfsits (no lr)	0.029	0.015	**14.7**	7.0	**5.1**	1.4
	tamrfsits (full)	**0.025**	0.013	18.2	6.9	4.7	1.5
B12	sen2like	0.036	0.020	65.8	12.5	0.2	2.0
	stair	0.025	0.013	52.5	5.1	0.8	1.6
	dh	0.029	0.016	64.1	5.0	-1.4	1.5
	dstfn	0.035	0.020	29.2	7.6	4.0	1.7
	tamrfsits (no hr)	0.021	0.011	37.9	5.3	3.5	1.5
	tamrfsits (no lr)	0.022	0.012	**15.5**	5.8	**5.6**	1.5
	tamrfsits (full)	0.**019**	0.010	18.1	6.2	5.3	1.4

Table 9: Comparison of the different methods on the spatio-temporal fusion task. Only a subset of bands is presented. Full results are available in appendix 5. * if usually higher FR means higher spatial resolution details, in the case of bands that are 10 meter native resolution, the best FR is the closest to zero in absolute value. Otherwise, ↓ (resp. ↑) indicates that the metric should be minimized (resp. maximized).Best mean values for each band and metric are highlighted in **bold**, and second best values are underlined.

with TAMRFSITS using only Landsat images. Looking a SWIR color compositions, TAMRFSITS is the only algorithm that provide sharpness and good IQ. DSTFN is sharper than the other competing algorithms, but bears artifacts similar to aliasing. On STAIR prediction, an artifact caused by temporal interpolation based on cloud masks can be observed around the larger of the two lakes. In conclusion, TAMRFSITS provides the best performances in this spatio-temporal task. It can use only the Landsat image and yield already consistent results, but the best performances are obtained by combining the Landsat image at target date with the SITS of Sentinel-2 images observed at other instants. Nevertheless, the same pre-trained TAMRFSITS model can do both, and even only use Sentinel-2 SITS for prediction. Another benefit of TAMRFSITS is that it handles seamlessly the case where the Landsat image at target date is cloudy, without using any additional mask. It is worth noting that none of the competing methods can make a prediction in this case, since they heavily rely on the LR observation at target date.

3.3.4. Thermal Sharpening

This last task focuses specifically on sharpening the LST band of Landsat, for observed Landsat dates. In this task, all available Sentinel-2 and Landsat dates are fed to TAMRFSITS, and the model is asked to predict the date of the Landsat images. We then measure specifically the performances of the LST band. As a baseline for comparison, the Data Mining Sharpener (DMS) Gao et al. (2012) is used. DMS is widely used in the Thermal Infra Red community, especially for operational productions. In these experiments, we used the pyDMS[2] implementation. In order to provide the High Resolution input to DMS, the closest Sentinel-2 image that meets a maximum cloud coverage criterion is selected. Only the 10 meters bands are given to DMS, and the default parameters of the method are used.

Table 10 shows the performances of TAMRFSITS and DMS on the thermal sharpening task, evaluated on the LS2S2 testing set. In terms of RMSE, DMS has better performances than TAMRFSITS. This is partly due to its residual compensation mech-

[2] https://github.com/radosuav/pyDMS

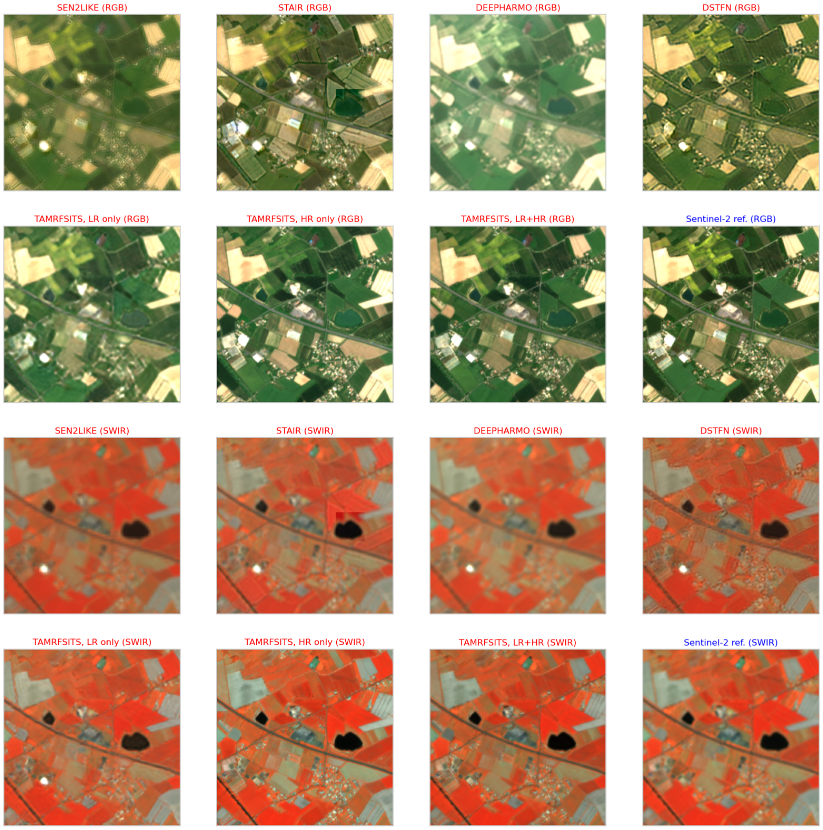


Figure 12: Predictions of the Spatio-Temporal Fusion task for all methods, on tile 31TCJ, 2022-05-10. First two rows shows the visible (RGB) color composition, and last two rows shows the SWIR (B8a, B11, B12) color composition.

anism, which ensures that the sharpened image matches exactly the input LST image downsampled back to its initial resolution. TAMRFSITS has a similar constraint expressed by the L_{LR}^{clear} loss term during training, but this constraint is not enforced at inference time. Still TAMRFSITS mean RMSE is around 1.2K, which is close to the expected accuracy of LST products. In terms of IQ, TAMRFSITS outputs are far better graded than DMS outputs. Spatial Frequency restoration is higher for DMS, but its very high BRISQUE score points out noisy reconstruction, which can explain higher FR.

	RMSE↓		**BRISQUE↓**		**FR↑**	
Method	mean	std	mean	std	mean	std
DMS	**0.024**	0.060	81.00	18.93	**4.2**	4.8
tamrfsits	1.291	1.133	**21.10**	24.28	3.2	2.3

Table 10: Comparison between DMS and TAMRFSITS on the Thermal Sharpening task. Best results for each metric are highlighted in **bold**. ↓ (resp. ↑) indicates that the metric should be minimized (resp. maximized).

This is confirmed by the visual inspection of sample predictions shown in Fig. 13: DMS predictions are noisier than TAMRFSITS predictions, while TAMRFSITS is slightly less accurate than DMS, predicting higher LST than what could be expected from the Landsat input LST. However, it must be stressed that TAMRFSITS is a versatile model that is able to sharpen LST among other capabilities, as opposed to ad hoc models that focuses on Thermal Sharpening only.

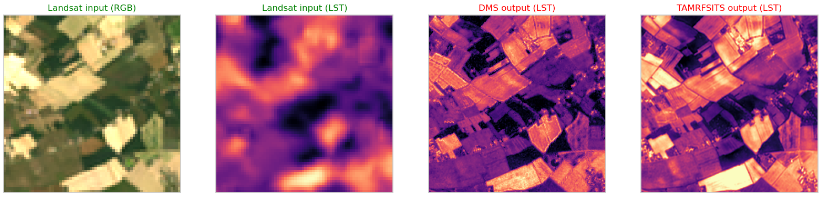


Figure 13: Predictions of the Thermal Sharpening task for DMS and TAMRFSITS, on tile 31TCJ, 2022-06-03. From left to right: Landsat RGB color composition, Landsat LST, Landsat LST sharpened by DMS, Landsat LST sharpened by TAMRFSITS.

3.4. Summary of experiments

In this section, a single pretrained TAMRFSITS model has been used to solve 4 different downstream tasks. For each of these tasks, the pre-trained model has been compared to algorithms from the literature which have been specifically designed to solve the task at stake. The experiments show that TAMRFSITS is on par or better than those algorithms in all the tasks, except for the Thermal Sharpening tasks where it still performs reasonably well with a better IQ. Those experiments also demonstrate the benefits of using a single versatile model: cloud removal when performing Sentinel-2 20-meter bands sharpening, or the ability to input Landsat SITS, Sentinel-2 SITS or both and always get consistent predictions without modifying and retraining the model. It is important to keep in mind that results presented in this section are measured in the frame of limitations imposed by the dataset as well as the competing algorithms. For instance, RMSE is always measured on pixels of acquired reference dates that are not masked, but TAMRFSITS also produce consistent outputs even when and where there is no reference value to compare to. In the spatio-temporal fusion task, only 6 Sentinel-2 bands are evaluated because the competing algorithms only process those 6 bands, but TAMRFSITS predicts the 10 Sentinel-2 bands and the 8 Landsat bands.

4. Discussions and Conclusion

4.1. Discussion

4.1.1. Training on larger datasets

An important limitation is the limited geographical and temporal span of the LS2S2 dataset. While it has been very useful to demonstrate the concept behind TAMRFSITS, it only spans a single year and has a geographical coverage mostly centered over Europe. If TAMRFSITS were to be used as a production model, it most certainly needs to be trained and evaluated on global scale, multi-year dataset.

4.1.2. Model complexity

In TAMRFSITS, the attention based operations are performed at the target resolution. Memory consumption of the temporal transformer is mostly linear with respect to the number of input dates, and working at target resolution means that this memory consumption is high, though each pixel can be processed separately at this stage. The number of input dates to the temporal Transfomer is also the main driver for inference time. Reducing the number of parameters of the transformer is the most straightforward action toward reducing inference cost and memory print. However, evaluating how this reduction affects the model performances is still to be determined.

4.1.3. Meteorological conditioning

Performances on Land Surface Temperature could be improved and extended to unseen dates by conditioning the prediction on meteorological data using data sources such as AGERA5. This would allow for a continuous prediction of LST through time, which is not currently possible.

4.1.4. Generalization to other sensors

TAMRFSITS is a very versatile model that solve many tasks without retraining. It supports variable number of input Sentinel-2 and Landsat dates and can predict a variable number of target dates. There is also nothing in the problem formulation or the model architecture that restrict TARMFSITS to only a pair of sensors. However, using different sensors or extending to a greater number of sensors will require adapting the model architecture, by including additional spatial encoders and extending the spatial decoder to predict more spectral bands. Then this new instance of the model will need to be trained for those new sensors. With the advent of foundation models, architectures such as PercieverIO Jaegle et al. (2021) could alleviate the need for those retraining by encoding the spectral, spatial and temporal context of each band for each sensor, yielding a generic model that could integrate any optical sensor, even if it has not been seen during training. This would require to revisit the architecture of TARMFSITS, but the training process proposed in this paper could be used to train such a model.

4.1.5. Toward new L3A products

TAMRFSITS is not trained for a specific task and can be used for various tasks, as demonstrated in section 3.3. One interesting way of using TAMRFSITS is generating joint Landsat and Sentinel-2 Level 3A monthly synthesis Hagolle et al. (2021). Those products are generated by weighted averaging of images comprised in each month, taking into account cloud masks. This level of product is very interesting for users, because it simplifies the use of the data by offering a regular, global time grid. However, because they rely on cloud masks to determine valid pixels, current methods can yield artifacts such as those shown in Fig. 9. The dates averaged for a given pixel also depend on the cloud mask, and each pixel is therefore a different mix of several dates. Current L3A products therefore do not correspond to a specific date.

TAMRFSITS can produce regular cloud-free L2 predictions with any time-step by using all observed Landsat and Sentinel-2 dates. Fig. 14 shows an example of such L2 products generated by TAMRFSITS, by requesting a prediction every 15th day of the month. Note that all bands, including B8A, B11 and B12 are true 10-meter resolution, and that there are no cloud mask interpolation artifacts. This could be used instead of L3 products. If one requires real monthly synthesis, the time-step could be reduced to 10 our 5 days and the predictions could be statistically aggregated. In both cases, the products for the user would be free of spurious cases and interpolation artifacts.

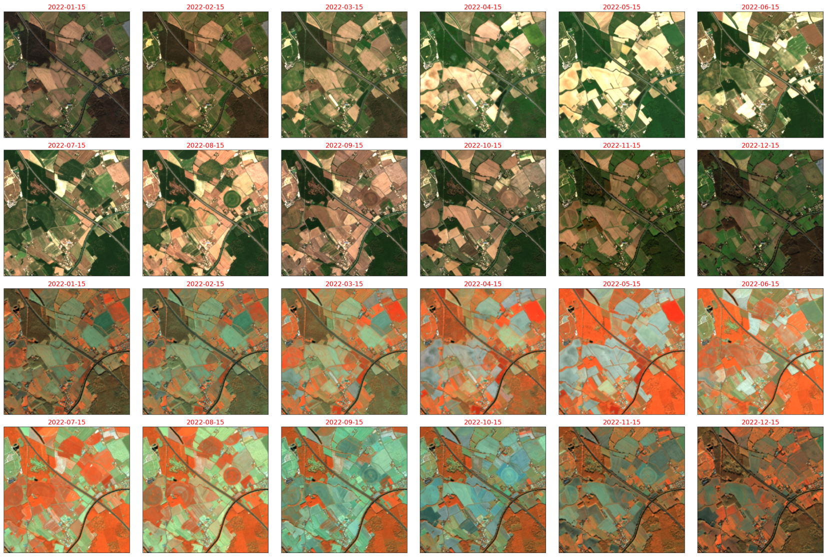


Figure 14: Level 3A monthly synthesis produced by TAMRFSITS from all observed Sentinel-2 and Landsat data over the 31TCJ AOIs. The two top rows displays the natural color composition (B4,B3, B2), while the two bottom rows display the SWIR color composition (B9A, B11, B12).

4.2. Conclusion

In this paper, we generalize the problem of temporal interpolation, spatial resolution enhancement and spatio-temporal fusion into a single original problem formulation. This problem formulation allows overcoming limitations that have been restricting the capabilities of methods proposed in the literature for decades. We then proposed TAMRFSITS, a DL architecture and training framework that allows to solve the generic problem formulation. This original architecture leverages a combination of Residual CNN for the spatial encoding with a Transformer for temporal encoding. Observation from both sensors are informed by Temporal Positional Encoding as well as a learnable sensor token, enabling the Transformer to exploit this information in the reconstruction. We demonstrated TAMRFSITS capabilities on a new dataset comprising one year of joint Sentinel-2 and Landsat SITS. The TAMRFSITS has unmatched versatility: it can process any number of Sentinel-2 dates and any number of Landsat dates, and it predicts all bands from both sensors at 10-meter spatial resolution, for any target date. All these capabilities are achieved by a single instance of the TAMRFSITS model, trained on a MAE pretext task. We compared TAMRFSITS to existing models in the literature on four tasks: gap-filling, band-sharpening of Sentinel-2 20-meter bands, Spatio-Temporal Fusion, and Thermal sharpening. For all these tasks, TAMRFSITS is on par or better than existing models, except for Thermal Sharpening, where it still shows good performances and high IQ. TAMRFSITS can also solve those tasks with capabilities that no other method can offer, such as cloud removal when performing Sentinel-2 20-meter bands sharpening, or the ability to input Landsat SITS, Sentinel-2 SITS or both and always get consistent predictions without modifying and retraining the model. Finally, we demonstrated the potential of TAMRFSITS in redefining Level 3A processing into a multi-sensor, spatial-resolution enhanced, temporally accurate and artifact free product. Our future work includes generalizing TAMRFSITS to any optical sensor and any spatial resolution, without retraining, which will result in a foundation model for SITS fusion. We also envision to condition TAMRFSITS predictions with exogenous data such as meteorological time-series. The complete source code for training and experiments is available here: https://github.com/Evoland-Land-Monitoring-Evolution/tamrfsits.

Acknowledgments

- This work was partly performed using HPC resources from GENCI-IDRIS (Grant 2023-AD010114835)
- This work was partly performed using HPC resources from CNES Computing Center.
- The authors acknowledge funding from the EvoLand project (Evolution of the Copernicus Land Service portfolio), grant agreement No 101082130) funded from the European Union's Horizon Europe research and innovation program.

CRediT

Julien Michel: Conceptualization, Formal Analysis, Methodology, Data Curation, Investigation, Software, Writing - Original Draft. **Jordi Inglada:** Conceptualization, Supervision, Writing - Review & Editing.

References

S. Bojinski, M. Verstraete, T. C. Peterson, C. Richter, A. Simmons, M. Zemp, The concept of essential climate variables in support of climate research, applications, and policy, Bulletin of the American Meteorological Society 95 (2014) 1431 – 1443. URL: https://journals.ametsoc.org/view/journals/bams/95/9/bams-d-13-00047.1.xml. doi:10.1175/BAMS-D-13-00047.1.

W. Jetz, M. A. McGeoch, R. Guralnick, S. Ferrier, J. Beck, M. J. Costello, M. Fernandez, G. N. Geller, P. Keil, C. Merow, et al., Essential biodiversity variables for mapping and monitoring species populations, Nature ecology & evolution 3 (2019) 539–551.

M. A. Wulder, D. P. Roy, V. C. Radeloff, T. R. Loveland, M. C. Anderson, D. M. Johnson, S. Healey, Z. Zhu, T. A. Scambos, N. Pahlevan, M. Hansen, N. Gorelick, C. J. Crawford, J. G. Masek, T. Hermosilla, J. C. White, A. S. Belward, C. Schaaf, C. E. Woodcock, J. L. Huntington, L. Lymburner, P. Hostert, F. Gao, A. Lyapustin, J.-F. Pekel, P. Strobl, B. D. Cook, Fifty years of landsat science and impacts, Remote Sensing of Environment 280 (2022) 113195. URL: http://dx.doi.org/10.1016/j.rse.2022.113195. doi:10.1016/j.rse.2022.113195.

G. Misra, F. Cawkwell, A. Wingler, Status of phenological research using Sentinel-2 data: A review, Remote Sensing 12 (2020) 2760.

D. Phiri, M. Simwanda, S. Salekin, V. R. Nyirenda, Y. Murayama, M. Ranagalage, Sentinel-2 data for land cover/use mapping: A review, Remote Sensing 12 (2020) 2291.

B. Vajsová, D. Fasbender, C. Wirnhardt, S. Lemajic, W. Devos, Assessing spatial limits of Sentinel-2 data on arable crops in the context of checks by monitoring, Remote Sensing 12 (2020) 2195. URL: http://dx.doi.org/10.3390/rs12142195. doi:10.3390/rs12142195.

U. Bhangale, S. More, T. Shaikh, S. Patil, N. More, Analysis of surface water resources using Sentinel-2 imagery, Procedia Computer Science 171 (2020) 2645–2654. URL: http://dx.doi.org/10.1016/j.procs.2020.04.287. doi:10.1016/j.procs.2020.04.287.

A. M. Wilson, W. Jetz, Remotely sensed high-resolution global cloud dynamics for predicting ecosystem and biodiversity distributions, PLOS Biology 14 (2016) e1002415. URL: http://dx.doi.org/10.1371/journal.pbio.1002415. doi:10.1371/journal.pbio.1002415.

J. Li, B. Chen, Global revisit interval analysis of landsat-8 -9 and Sentinel-2a -2b data for terrestrial monitoring, Sensors 20 (2020) 6631. URL: http://dx.doi.org/10.3390/s20226631. doi:10.3390/s20226631.

N. Latte, P. Lejeune, Planetscope radiometric normalization and Sentinel-2 super-resolution (2.5 m): A straightforward spectral-spatial fusion of multi-satellite multi-sensor images using residual convolutional neural networks, Remote Sensing 12 (2020) 2366.

Y. Sadeh, X. Zhu, D. Dunkerley, J. P. Walker, Y. Zhang, O. Rozenstein, V. Manivasagam, K. Chenu, Fusion of Sentinel-2 and planetscope time-series data into daily 3 m surface reflectance and wheat lai monitoring, International Journal of Applied Earth Observation and Geoinformation 96 (2021) 102260.

J.-P. Lagouarde, B. Bhattacharya, P. Crébassol, P. Gamet, S. S. Babu, G. Boulet, X. Briottet, K. Buddhiraju, S. Cherchali, I. Dadou, G. Dedieu, M. Gouhier, O. Hagolle, M. Irvine, F. Jacob, A. Kumar, K. K. Kumar, B. Laignel, K. Mallick, C. Murthy, A. Olioso, C. Ottlé, M. R. Pandya, P. V. Raju, J.-L. Roujean, M. Sekhar, M. V. Shukla, S. K. Singh, J. Sobrino, R. Ramakrishnan, The Indian-French Trishna Mission: Earth Observation in the Thermal Infrared with High Spatio-Temporal Resolution, in: IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium, 2018, pp. 4078–4081. doi:10.1109/IGARSS.2018.8518720, iSSN: 2153-7003.

F. Bernard, I. Manolis, I. Barat, A. B. Alamañac, M. S. Taboada, P. Mingorance, A. Ciapponi, T. Cardone, I. F. Nunez, A. Garcia, P. Hallibert, A. Hammar, P. Henriot, D. M. Codinachs, S. Patti, P. P. Montes, P. Skrzypek, D. Steenari, S. Weixler, S. Deslous, C. Coatantiec, O. A. Trotta, I. C. Vega, D. G. Holgueras, The copernicus land surface temperature monitoring (lstm) mission: design, technology and status, in: Sensors, Systems, and Next-Generation Satellites XXVII, 2023, p. 8. URL: http://dx.doi.org/10.1117/12.2679705. doi:10.1117/12.2679705.

S. Anwar, S. Khan, N. Barnes, A deep journey into super-resolution: A survey, ACM Computing Surveys (CSUR) 53 (2020) 1–34.

A. Liu, Y. Liu, J. Gu, Y. Qiao, C. Dong, Blind image super-resolution: a survey and beyond, IEEE Transactions on Pattern Analysis and Machine Intelligence (2022) 1–19. URL: http://dx.doi.org/10.1109/TPAMI.2022.3203009. doi:10.1109/tpami.2022.3203009.

X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, C. Change Loy, ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks, in: Proceedings of the European Conference on Computer Vision (ECCV) Workshops, 2018.

C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, W. Shi, Photo-realistic single image super-resolution using a generative adversarial network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

M. Galar, R. Sesma, C. Ayala, C. Aranda, Super-resolution for Sentinel-2 images, International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences (2019).

N. L. Nguyen, J. Anger, L. Raad, B. Galerne, G. Facciolo, On The Role of Alias and Band-Shift for Sentinel-2 Super-Resolution, in: IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium, 2023, pp. 4294–4297. doi:10.1109/IGARSS52108.2023.10282805.

L. Salgueiro Romero, J. Marcello, V. Vilaplana, Super-resolution of Sentinel-2 imagery using generative adversarial networks, Remote Sensing 12 (2020) 2424.

D. Pouliot, R. Latifovic, J. Pasher, J. Duffe, Landsat super-resolution enhancement using convolution neural networks and Sentinel-2 for training, Remote Sensing 10 (2018) 394.

M. Märtens, D. Izzo, A. Krzic, D. Cox, Super-resolution of PROBA-V images using convolutional neural networks, Astrodynamics 3 (2019) 387–402.

A. Okabayashi, N. Audebert, S. Donike, C. Pelletier, Cross-sensor super-resolution of irregularly sampled Sentinel-2 time series, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2024, pp. 502–511.

A. B. Molini, D. Valsesia, G. Fracastoro, E. Magli, Deepsum: Deep neural network for super-resolution of unregistered multitemporal images, IEEE Transactions on Geoscience and Remote Sensing 58 (2019) 3644–3656.

M. R. Ibrahim, R. Benavente, D. Ponsa, F. Lumbreras, Hyda-net: a hybrid dense attention network for remote sensing multi-image super-resolution, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (2025) 1–24. URL:

http://dx.doi.org/10.1109/JSTARS.2025.3547785. doi:10.1109/jstars.2025.3547785.

J. Michel, E. Kalinicheva, J. Inglada, Revisiting remote sensing cross-sensor single image super-resolution: the overlooked impact of geometric and radiometric distortion, IEEE Transactions on Geoscience and Remote Sensing (2025) 1–1. doi:10.1109/TGRS.2025.3572548.

V. T. Sambandham, K. Kirchheim, F. Ortmeier, S. Mukhopadhaya, Deep learning-based harmonization and super-resolution of landsat-8 and sentinel-2 images, ISPRS Journal of Photogrammetry and Remote Sensing 212 (2024) 274–288. URL: http://dx.doi.org/10.1016/j.isprsjprs.2024.04.026. doi:10.1016/j.isprsjprs.2024.04.026.

M. K. Firozjaei, M. Kiavarz, S. K. Alavipanah, Satellite-derived land surface temperature spatial sharpening: a comprehensive review on current status and perspectives, European Journal of Remote Sensing 55 (2022) 644–664. URL: http://dx.doi.org/10.1080/22797254.2022.2144764. doi:10.1080/22797254.2022.2144764.

G. Vivone, L. Alparone, J. Chanussot, M. Dalla Mura, A. Garzelli, G. A. Licciardi, R. Restaino, L. Wald, A critical comparison among pansharpening algorithms, IEEE Transactions on Geoscience and Remote Sensing 53 (2014) 2565–2586.

M. Ciotola, G. Guarino, G. Vivone, G. Poggi, J. Chanussot, A. Plaza, G. Scarpa, Hyperspectral pansharpening: Critical review, tools, and future perspectives, IEEE Geoscience and Remote Sensing Magazine (2025) 2–29. URL: http://dx.doi.org/10.1109/MGRS.2024.3509139. doi:10.1109/mgrs.2024.3509139.

H. Su, Y. Li, Y. Xu, X. Fu, S. Liu, A review of deep-learning-based super-resolution: From methods to applications, Pattern Recognition 157 (2025) 110935. URL: http://dx.doi.org/10.1016/j.patcog.2024.110935. doi:10.1016/j.patcog.2024.110935.

R. d. Lutio, S. D'aronco, J. D. Wegner, K. Schindler, Guided super-resolution as pixel-to-pixel transformation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 8829–8837.

L. Wald, T. Ranchin, M. Mangolini, Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images, Photogrammetric engineering and remote sensing 63 (1997) 691–699.

F. Palsson, J. R. Sveinsson, M. O. Ulfarsson, Sentinel-2 image fusion using a deep residual network, Remote Sensing 10 (2018). URL: https://www.mdpi.com/2072-4292/10/8/1290. doi:10.3390/rs10081290.

F. Gao, W. P. Kustas, M. C. Anderson, A data mining approach for sharpening thermal satellite imagery over land, Remote Sensing 4 (2012) 3287–3319. URL: https://www.mdpi.com/2072-4292/4/11/3287. doi:10.3390/rs4113287.

C. Granero-Belinchon, A. Michel, J.-P. Lagouarde, J. A. Sobrino, X. Briottet, Multi-resolution study of thermal unmixing techniques over madrid urban area: Case study of trishna mission, Remote Sensing 11 (2019). URL: https://www.mdpi.com/2072-4292/11/10/1251. doi:10.3390/rs11101251.

L. Salgueiro, J. Marcello, V. Vilaplana, Single-image super-resolution of sentinel-2 low resolution bands with residual dense convolutional neural networks, Remote Sensing 13 (2021) 5007. URL: http://dx.doi.org/10.3390/rs13245007. doi:10.3390/rs13245007.

B. M. Nguyen, G. Tian, M.-T. Vo, A. Michel, T. Corpetti, C. Granero-Belinchon, Convolutional neural network modelling for modis land surface temperature super-resolution, in: 2022 30th European Signal Processing Conference (EUSIPCO), 2022, pp. 1806–1810. URL: http://dx.doi.org/10.23919/EUSIPCO55093.2022.9909569. doi:10.23919/eusipco55093.2022.9909569.

O. Merlin, B. Duchemin, O. Hagolle, F. Jacob, B. Coudert, G. Chehbouni, G. Dedieu, J. Garatuza, Y. Kerr, Disaggregation of modis surface temperature over an agricultural area using a time series of formosat-2 images, Remote Sensing of Environment

114 (2010) 2500–2512. URL: http://dx.doi.org/10.1016/j.rse.2010.05.025. doi:10.1016/j.rse.2010.05.025.

Z. Shao, J. Cai, P. Fu, L. Hu, T. Liu, Deep learning-based fusion of landsat-8 and Sentinel-2 images for a harmonized surface reflectance product, Remote Sensing of Environment 235 (2019) 111425. URL: http://dx.doi.org/10.1016/j.rse.2019.111425. doi:10.1016/j.rse.2019.111425.

M. Belgiu, A. Stein, Spatiotemporal image fusion in remote sensing, Remote sensing 11 (2019) 818.

J. Xiao, A. K. Aggarwal, N. H. Duc, A. Arya, U. K. Rage, R. Avtar, A review of remote sensing image spatiotemporal fusion: Challenges, applications and recent trends, Remote Sensing Applications: Society and Environment 32 (2023) 101005. URL: http://dx.doi.org/10.1016/j.rsase.2023.101005. doi:10.1016/j.rsase.2023.101005.

F. Gao, J. Masek, M. Schwaller, F. Hall, On the blending of the landsat and modis surface reflectance: Predicting daily landsat surface reflectance, IEEE Transactions on Geoscience and Remote sensing 44 (2006) 2207–2218.

X. Zhu, F. Cai, J. Tian, T. K.-A. Williams, Spatiotemporal fusion of multisource remote sensing data: Literature survey, taxonomy, principles, applications, and future directions, Remote Sensing 10 (2018) 527.

Z. Tan, P. Yue, L. Di, J. Tang, Deriving high spatiotemporal remote sensing images using deep convolutional network, Remote Sensing 10 (2018) 1066. URL: http://dx.doi.org/10.3390/rs10071066. doi:10.3390/rs10071066.

Z. Tan, L. Di, M. Zhang, L. Guo, M. Gao, An enhanced deep convolutional model for spatiotemporal image fusion, Remote Sensing 11 (2019) 2898. URL: http://dx.doi.org/10.3390/rs11242898. doi:10.3390/rs11242898.

X. Liu, C. Deng, J. Chanussot, D. Hong, B. Zhao, <i>stfnet</i>: a two-stream convolutional neural network for spatiotemporal image fusion, IEEE Transactions on

Geoscience and Remote Sensing 57 (2019) 6552–6564. URL: http://dx.doi.org/10.1109/TGRS.2019.2907310. doi:10.1109/tgrs.2019.2907310.

Z. Tan, M. Gao, X. Li, L. Jiang, A flexible reference-insensitive spatiotemporal fusion model for remote sensing images using conditional generative adversarial network, IEEE Transactions on Geoscience and Remote Sensing 60 (2022) 1–13. doi:10.1109/TGRS.2021.3050551.

Y. Zhang, R. Fan, P. Duan, J. Dong, Z. Lei, Dcdgan-stf: a multiscale deformable convolution distillation gan for remote sensing image spatiotemporal fusion (2024) 1–15. URL: http://dx.doi.org/10.1109/JSTARS.2024.3476153. doi:10.1109/jstars.2024.3476153.

Y. Xie, J. Hu, K. He, L. Cao, K. Yang, L. Chen, The gan spatiotemporal fusion model based on multi-scale convolution and attention mechanism for remote sensing images, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (2024) 1–13. URL: http://dx.doi.org/10.1109/JSTARS.2024.3507815. doi:10.1109/jstars.2024.3507815.

Y. Wu, M. Huang, A unified generative adversarial network with convolution and transformer for remote sensing image fusion, IEEE Transactions on Geoscience and Remote Sensing (2024) 1–1. URL: http://dx.doi.org/10.1109/TGRS.2024.3441719. doi:10.1109/tgrs.2024.3441719.

G. Yang, Y. Qian, H. Liu, B. Tang, R. Qi, Y. Lu, J. Geng, Msfusion: Multistage for remote sensing image spatiotemporal fusion based on texture transformer and convolutional neural network, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 15 (2022) 4653–4666. URL: http://dx.doi.org/10.1109/JSTARS.2022.3179415. doi:10.1109/jstars.2022.3179415.

W. Li, D. Cao, M. Xiang, Enhanced multi-stream remote sensing spatiotemporal fusion network based on transformer and dilated convolution, Remote Sensing 14 (2022) 4544. URL: http://dx.doi.org/10.3390/rs14184544. doi:10.3390/rs14184544.

G. Chen, P. Jiao, Q. Hu, L. Xiao, Z. Ye, Swinstfm: Remote sensing spatiotemporal fusion using swin transformer, IEEE Transactions on Geoscience and Remote Sensing 60 (2022) 1–18. URL: http://dx.doi.org/10.1109/TGRS.2022.3182809. doi:10.1109/tgrs.2022.3182809.

H. Goyena, U. Pérez-Goya, M. Montesino-SanMartin, A. F. Militino, Q. Wang, P. M. Atkinson, M. D. Ugarte, Unpaired spatio-temporal fusion of image patches (ustfip) from cloud covered images, Remote Sensing of Environment 295 (2023) 113709. URL: http://dx.doi.org/10.1016/j.rse.2023.113709. doi:10.1016/j.rse.2023.113709.

X. Zhang, S. Li, Z. Tan, X. Li, Enhanced wavelet based spatiotemporal fusion networks using cross-paired remote sensing images, ISPRS Journal of Photogrammetry and Remote Sensing 211 (2024) 281–297. URL: http://dx.doi.org/10.1016/j.isprsjprs.2024.04.016. doi:10.1016/j.isprsjprs.2024.04.016.

W. R. Moskolaï, W. Abdou, A. Dipanda, Kolyang, Application of deep learning architectures for satellite image time series prediction: a review, Remote Sensing 13 (2021) 4822. URL: http://dx.doi.org/10.3390/rs13234822. doi:10.3390/rs13234822.

S. Li, L. Xu, Y. Jing, H. Yin, X. Li, X. Guan, High-quality vegetation index product generation: a review of ndvi time series reconstruction techniques, International Journal of Applied Earth Observation and Geoinformation 105 (2021) 102640. URL: http://dx.doi.org/10.1016/j.jag.2021.102640. doi:10.1016/j.jag.2021.102640.

D. P. Roy, J. Ju, P. Lewis, C. Schaaf, F. Gao, M. Hansen, E. Lindquist, Multi-temporal modis–landsat data fusion for relative radiometric normalization, gap filling, and prediction of landsat data, Remote Sensing of Environment 112 (2008) 3112–3130.

B. Irigireddy, V. Bandaru, Satflow: Generative model based framework for producing high resolution gap free remote sensing imagery, 2025. URL: https://arxiv.org/abs/2502.01098. arXiv:2502.01098.

H. Liu, H. K. Zhang, B. Huang, L. Yan, K. K. Tran, Y. Qiu, X. Zhang, D. P. Roy, Reconstruction of seamless harmonized landsat sentinel-2 (hls) time series via self-supervised learning, Remote Sensing of Environment 308 (2024) 114191. URL: http://dx.doi.org/10.1016/j.rse.2024.114191. doi:10.1016/j.rse.2024.114191.

M. Claverie, J. Ju, J. G. Masek, J. L. Dungan, E. F. Vermote, J.-C. Roger, S. V. Skakun, C. Justice, The harmonized landsat and Sentinel-2 surface reflectance data set, Remote Sensing of Environment 219 (2018) 145 – 161. doi:https://doi.org/10.1016/j.rse.2018.09.002.

S. Saunier, B. Pflug, I. Lobos, B. Franch, J. Louis, R. D. L. Reyes, V. Debaecker, E. Cadau, V. Boccia, F. Gascon, S. Kocaman, Sen2like: Paving the way towards harmonization and fusion of optical data, Remote Sensing 14 (2022) 3855. URL: http://dx.doi.org/10.3390/rs14163855. doi:10.3390/rs14163855.

Q. Wang, G. A. Blackburn, A. O. Onojeghuo, J. Dash, L. Zhou, Y. Zhang, P. M. Atkinson, Fusion of landsat 8 oli and Sentinel-2 msi data, IEEE Transactions on Geoscience and Remote Sensing 55 (2017) 3885–3899. URL: http://dx.doi.org/10.1109/TGRS.2017.2683444. doi:10.1109/tgrs.2017.2683444.

C. Dong, C. C. Loy, K. He, X. Tang, Image super-resolution using deep convolutional networks, IEEE transactions on pattern analysis and machine intelligence 38 (2015) 295–307.

Y. Luo, K. Guan, J. Peng, Stair: A generic and fully-automated method to fuse multiple sources of optical satellite data to generate a high-resolution, daily and cloud-/gap-free surface reflectance product, Remote Sensing of Environment 214 (2018) 87–99.

Y. Luo, K. Guan, J. Peng, S. Wang, Y. Huang, Stair 2.0: A generic and automatic algorithm to fuse modis, landsat, and Sentinel-2 to generate 10 m, daily, and cloud-/gap-free surface reflectance product, Remote Sensing 12 (2020) 3209.

P. Wang, M. Huang, S. Shi, B. Huang, B. Zhou, G. Xu, L. Wang, H. Leung, Landsat-8 and Sentinel-2 image fusion based on multi-scale smoothing-sharpening filter, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (2024) 1–14. URL: http://dx.doi.org/10.1109/JSTARS.2024.3469974. doi:10.1109/jstars.2024.3469974.

R. Mukherjee, D. Liu, Spatial and spectral translation of landsat 8 to sentinel-2 using conditional generative adversarial networks, Remote Sensing 15 (2023) 5502. URL: http://dx.doi.org/10.3390/rs15235502. doi:10.3390/rs15235502.

B. Chen, J. Li, Y. Jin, Deep learning for feature-level data fusion: Higher resolution reconstruction of historical landsat archive, Remote Sensing 13 (2021) 167. URL: http://dx.doi.org/10.3390/rs13020167. doi:10.3390/rs13020167.

J. Chang, W. Du, B. Zhang, S. Guo, Y. Yin, Z. Wang, T. Xu, Z. Feng, Based on the improved edcstfn model, modis, landsat 8 and sentinel-2 data were fused to obtain 10m dense time series images, IEEE Access (2025) 1–1. URL: http://dx.doi.org/10.1109/ACCESS.2025.3564968. doi:10.1109/access.2025.3564968.

I. Dumeur, S. Valero, J. Inglada, Self-supervised spatio-temporal representation learning of satellite image time series, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 17 (2024) 4350–4367. URL: http://dx.doi.org/10.1109/JSTARS.2024.3358066. doi:10.1109/jstars.2024.3358066.

Y. Yuan, L. Lin, Q. Liu, R. Hang, Z.-G. Zhou, Sits-former: a pre-trained spatio-spectral-temporal representation model for sentinel-2 time series classification, International Journal of Applied Earth Observation and Geoinformation 106 (2022) 102651. URL: http://dx.doi.org/10.1016/j.jag.2021.102651. doi:10.1016/j.jag.2021.102651.

X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, C. Change Loy, Esrgan: Enhanced super-resolution generative adversarial networks, in: Proceedings of the European conference on computer vision (ECCV) workshops, 2018, pp. 0–0.

A. Vaswani, Attention is all you need, Advances in Neural Information Processing Systems (2017).

X. Guo, J. Lao, B. Dang, Y. Zhang, L. Yu, L. Ru, L. Zhong, Z. Huang, K. Wu, D. Hu, H. He, J. Wang, J. Chen, M. Yang, Y. Zhang, Y. Li, Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27672–27683.

I. Dumeur, S. Valero, J. Inglada, Paving the way toward foundation models for irregular and unaligned satellite image time series, 2024. URL: https://arxiv.org/abs/2407.08448. arXiv:2407.08448.

K. He, X. Chen, S. Xie, Y. Li, P. Dollár, R. Girshick, Masked autoencoders are scalable vision learners, 2021. arXiv:2111.06377.

H. Liu, J. Gan, X. Fan, Y. Zhang, C. Luo, J. Zhang, G. Jiang, Y. Qian, C. Zhao, H. Ma, Z. Guo, Pt-tuning: Bridging the gap between time series masked reconstruction and forecasting via prompt token tuning, nil nil (2023) nil. URL: https://arxiv.org/abs/2311.03768. doi:10.48550/ARXIV.2311.03768.

C. J. Reed, R. Gupta, S. Li, S. Brockman, C. Funk, B. Clipp, K. Keutzer, S. Candido, M. Uyttendaele, T. Darrell, Scale-mae: a scale-aware masked autoencoder for multiscale geospatial representation learning, CoRR (2022). URL: http://arxiv.org/abs/2212.14532v4. arXiv:2212.14532v4.

R. Girshick, Fast R-CNN, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015.

V. Balntas, E. Riba, D. Ponsa, K. Mikolajczyk, Learning local feature descriptors with triplets and shallow convolutional neural networks, in: Procedings of the British Machine Vision Conference 2016, 2016, pp. 119.1–119.11. URL: http://dx.doi.org/10.5244/C.30.119. doi:10.5244/c.30.119.

R. Zhang, P. Isola, A. A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

M. Schramm, E. Pebesma, M. Milenković, L. Foresta, J. Dries, A. Jacob, W. Wagner, M. Mohr, M. Neteler, M. Kadunc, T. Miksa, P. Kempeneers, J. Verbesselt, B. Gößwein, C. Navacchi, S. Lippens, J. Reiche, The openeo api-harmonising the use of earth observation cloud services using virtual data cube functionalities, Remote Sensing 13 (2021) 1125. URL: http://dx.doi.org/10.3390/rs13061125. doi:10.3390/rs13061125.

D. P. Kingma, J. Ba, Adam: a Method for Stochastic Optimization (2014). URL: https://arxiv.org/abs/1412.6980. doi:10.48550/ARXIV.1412.6980.

I. Loshchilov, F. Hutter, SGDR: Stochastic Gradient Descent With Warm Restarts (2016). URL: https://arxiv.org/abs/1608.03983. doi:10.48550/ARXIV.1608.03983.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, Pytorch: An imperative style, high-performance deep learning library, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems 32, Curran Associates, Inc., 2019, pp. 8024–8035. URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

A. Mittal, A. K. Moorthy, A. C. Bovik, No-reference image quality assessment in the spatial domain, IEEE Transactions on Image Processing 21 (2012) 4695–4708. doi:10.1109/TIP.2012.2214050.

V. Lonjou, C. Desjardins, O. Hagolle, B. Petrucci, T. Tremas, M. Dejus, A. Makarau, S. Auer, Maccs-atcor joint algorithm (MAJA), in: Remote Sensing of Clouds and the Atmosphere XXI, volume 10001, International Society for Optics and Photonics, 2016, p. 1000107.

K. Li, K. Guan, C. Jiang, S. Wang, B. Peng, Y. Cai, Evaluation of four new land surface temperature (lst) products in the us corn belt: Ecostress, goes-r, landsat, and sentinel-3, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 14 (2021) 9931–9945.

C. Lanaras, J. Bioucas-Dias, S. Galliani, E. Baltsavias, K. Schindler, Super-resolution of Sentinel-2 images: Learning a globally applicable deep neural network, ISPRS Journal of Photogrammetry and Remote Sensing 146 (2018) 305–319.

J. Wu, L. Lin, T. Li, Q. Cheng, C. Zhang, H. Shen, Fusing landsat 8 and sentinel-2 data for 10-m dense time-series imagery using a degradation-term constrained deep network, International Journal of Applied Earth Observation and Geoinformation 108 (2022) 102738. URL: http://dx.doi.org/10.1016/j.jag.2022.102738. doi:10.1016/j.jag.2022.102738.

A. Jaegle, S. Borgeaud, J.-B. Alayrac, C. Doersch, C. Ionescu, D. Ding, S. Koppula, D. Zoran, A. Brock, E. Shelhamer, O. Hénaff, M. M. Botvinick, A. Zisserman, O. Vinyals, J. Carreira, Perceiver io: a general architecture for structured inputs & outputs (2021). URL: https://arxiv.org/abs/2107.14795. doi:10.48550/ARXIV.2107.14795.

O. Hagolle, J. Colin, S. Coustance, P. Kettig, P. D'Angelo, S. Auer, G. Doxani, C. Desjardins, Sentinel-2 surface reflectance products generated by cnes and dlr: Methods, validation and applications, ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences V-1-2021 (2021) 9–15. URL: http://dx.doi.org/10.5194/isprs-annals-V-1-2021-9-2021. doi:10.5194/isprs-annals-v-1-2021-9-2021.

5. Appendix: Complete results

		RMSE↓				**BRISQUE↓**				**FR↑**			
		Clear		Masked		Clear		Masked		Clear		Masked	
Band	**Method**	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
LS B1	naive	0.001	0.000	0.008	0.004	87.7	6.8	87.7	7.3	-0.0	1.2	-0.2	1.4
	tamrfsits	0.005	0.016	0.009	0.007	45.6	13.9	43.4	19.7	6.8	1.8	7.3	2.3
LS B2	naive	0.002	0.001	0.009	0.004	83.7	7.1	83.7	7.6	-0.0	1.2	-0.2	1.2
	tamrfsits	0.005	0.015	0.009	0.008	40.2	12.8	36.0	17.7	7.1	1.8	7.5	2.2
LS B3	naive	0.003	0.001	0.012	0.006	76.4	6.7	76.8	7.5	0.0	0.9	-0.0	0.9
	tamrfsits	0.006	0.014	0.011	0.008	32.2	9.7	23.6	10.0	7.2	1.8	7.8	2.0
LS B4	naive	0.004	0.001	0.017	0.008	70.3	6.1	71.0	7.2	0.1	0.8	-0.2	1.0
	tamrfsits	0.007	0.012	0.013	0.011	29.7	10.0	19.3	9.5	7.3	1.7	7.8	1.9
LS B5	naive	0.007	0.003	0.035	0.020	63.9	4.0	63.4	2.6	-0.0	0.7	0.0	1.0
	tamrfsits	0.011	0.014	0.025	0.014	31.6	11.1	12.5	6.9	6.8	1.9	7.8	2.0
LS B6	naive	0.006	0.001	0.031	0.016	64.2	3.5	64.2	3.1	0.1	0.9	-0.1	0.9
	tamrfsits	0.009	0.007	0.023	0.013	30.6	11.8	12.5	6.3	7.1	1.7	8.0	1.9
LS B7	naive	0.005	0.001	0.028	0.016	66.4	4.4	66.5	4.4	0.1	0.9	-0.1	0.9
	tamrfsits	0.008	0.008	0.019	0.013	29.8	11.3	13.3	6.7	7.1	1.6	7.9	1.8
LS LST	naive	0.440	0.425	4.290	3.196	73.0	10.1	76.2	12.8	-0.4	2.1	-0.0	2.1
	tamrfsits	1.169	0.659	13.060	8.325	22.1	22.0	68.8	23.9	3.1	2.4	18.2	6.1
S2 B2	naive	0.000	0.000	0.028	0.054	23.4	19.0	25.6	20.2	1.3	3.0	1.2	3.1
	tamrfsits	0.009	0.016	0.026	0.049	19.9	10.2	26.3	11.6	1.7	3.3	1.5	3.7
S2 B3	naive	0.000	0.000	0.029	0.052	19.6	17.3	21.1	18.3	1.2	2.8	1.1	2.8
	tamrfsits	0.009	0.015	0.027	0.048	15.0	6.2	20.2	6.8	1.5	3.0	1.3	3.4
S2 B4	naive	0.000	0.000	0.031	0.052	22.6	18.4	24.0	19.4	1.1	2.7	1.0	2.7
	tamrfsits	0.010	0.015	0.029	0.048	17.1	7.3	20.8	8.0	1.4	2.8	1.2	3.1
S2 B5	naive	0.006	0.005	0.032	0.051	49.7	11.6	50.3	12.0	1.2	2.3	1.1	2.1
	tamrfsits	0.011	0.015	0.030	0.046	12.5	5.8	15.1	5.6	6.0	2.2	5.9	2.6
S2 B6	naive	0.008	0.005	0.040	0.046	48.6	11.4	49.3	11.9	1.1	2.2	1.0	2.2
	tamrfsits	0.013	0.015	0.037	0.040	12.6	6.0	14.6	6.3	5.9	1.9	5.7	2.4
S2 B7	naive	0.009	0.005	0.043	0.043	47.9	11.5	48.6	12.1	1.1	2.1	0.9	2.1
	tamrfsits	0.014	0.015	0.039	0.038	13.5	6.0	14.5	6.5	5.6	1.8	5.6	2.2
S2 B8	naive	0.000	0.000	0.046	0.044	16.5	17.2	17.4	18.2	0.9	2.2	0.8	2.4
	tamrfsits	0.014	0.016	0.043	0.038	12.0	6.3	15.8	7.0	0.8	2.0	0.7	2.4
S2 B8a	naive	0.009	0.004	0.043	0.040	48.6	11.5	49.3	12.0	1.0	2.1	0.9	2.1
	tamrfsits	0.014	0.015	0.040	0.035	13.7	6.3	14.6	6.9	5.6	1.7	5.6	2.2
S2 B11	naive	0.005	0.002	0.031	0.021	55.9	11.1	56.3	11.4	1.0	2.0	1.0	2.0
	tamrfsits	0.011	0.007	0.030	0.019	14.4	10.0	14.3	6.4	6.6	1.8	6.2	2.3
S2 B12	naive	0.005	0.002	0.025	0.018	55.4	11.8	56.0	12.2	1.0	2.0	1.0	2.0
	tamrfsits	0.010	0.006	0.024	0.017	15.5	9.6	15.3	6.4	6.4	1.9	6.1	2.3

Table 11: Comparison between TAMRFSITS model and naive interpolation on the gap-filling task, where regular gaps of 30 days are masked from the input SITS and kept appart for validation. Only a selection of spectral bands is presented. Full results are available in appendix 5. ↓ (resp. ↑) indicates that the metric should be minimized (resp. maximized). Best mean values for each metric and each band are highlithed in **bold**.

		RMSE↓				**BRISQUE↓**				**FR↑**			
		Clear		Masked		Clear		Masked		Clear		Masked	
Band	**Method**	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
LS B1	naive	0.002	0.001	0.010	0.008	84.6	6.1	84.8	6.0	0.9	1.9	0.7	2.0
	tamrfsits	0.005	0.018	0.008	0.004	42.8	14.7	46.1	19.8	6.9	1.8	7.2	2.1
LS B2	naive	0.002	0.001	0.011	0.010	81.8	5.6	81.9	5.4	0.9	2.0	0.7	2.1
	tamrfsits	0.005	0.018	0.008	0.005	37.5	13.8	38.4	17.3	7.2	1.8	7.4	2.1
LS B3	naive	0.003	0.001	0.015	0.014	76.9	4.4	76.9	4.1	1.1	2.3	0.8	2.4
	tamrfsits	0.006	0.017	0.010	0.007	29.0	10.4	24.9	9.3	7.3	1.8	7.6	2.0
LS B4	naive	0.005	0.002	0.020	0.018	72.6	3.7	72.6	3.4	0.9	1.9	0.9	2.3
	tamrfsits	0.007	0.018	0.014	0.010	26.6	10.9	21.2	9.0	7.4	1.6	7.6	1.9
LS B5	naive	0.006	0.001	0.031	0.025	71.1	2.5	71.6	2.1	1.4	2.8	1.0	2.8
	tamrfsits	0.012	0.012	0.027	0.014	27.5	12.2	13.6	7.2	7.0	1.9	7.7	2.0
LS B6	naive	0.007	0.001	0.035	0.028	69.0	1.4	69.2	1.4	1.1	2.4	1.0	2.4
	tamrfsits	0.010	0.012	0.024	0.014	26.4	13.3	13.2	5.6	7.4	1.7	7.7	1.8
LS B7	naive	0.006	0.001	0.030	0.027	69.9	2.0	70.0	1.8	1.0	2.0	0.9	2.2
	tamrfsits	0.009	0.013	0.020	0.014	25.6	12.8	14.7	6.0	7.3	1.5	7.7	1.7
LS LST	naive	0.460	0.449			82.3	9.8			1.9	4.5		
	tamrfsits	1.270	0.927			19.3	20.0			3.2	2.1		
S2 B2	naive	0.000	0.000	0.031	0.055	18.9	8.3	20.9	8.9	1.8	3.3	1.7	3.8
	tamrfsits	0.010	0.019	0.028	0.051	20.6	10.6	26.5	11.3	1.7	3.3	1.5	3.9
S2 B3	naive	0.000	0.000	0.032	0.055	14.9	5.4	16.6	5.6	1.6	3.0	1.4	3.4
	tamrfsits	0.010	0.018	0.029	0.050	15.1	6.4	20.6	6.7	1.5	3.1	1.3	3.5
S2 B4	naive	0.000	0.000	0.036	0.055	17.5	6.8	18.5	6.7	1.4	2.8	1.2	3.1
	tamrfsits	0.011	0.018	0.032	0.049	17.5	8.0	20.9	7.8	1.4	2.8	1.2	3.3
S2 B5	naive	0.006	0.005	0.037	0.053	47.3	4.5	47.8	4.8	1.7	2.5	1.5	2.6
	tamrfsits	0.012	0.018	0.033	0.048	12.3	6.0	15.3	5.7	6.0	2.3	5.8	2.6
S2 B6	naive	0.008	0.004	0.045	0.046	47.2	4.5	47.9	4.9	1.4	2.2	1.3	2.4
	tamrfsits	0.014	0.017	0.040	0.042	12.5	6.2	14.5	6.2	5.8	1.9	5.7	2.4
S2 B7	naive	0.009	0.004	0.048	0.042	46.5	4.6	47.1	4.8	1.3	2.1	1.2	2.3
	tamrfsits	0.015	0.017	0.043	0.039	13.4	6.2	14.3	6.4	5.5	1.7	5.5	2.2
S2 B8	naive	0.000	0.000	0.052	0.043	12.9	6.3	13.9	6.0	0.8	2.0	0.7	2.4
	tamrfsits	0.015	0.017	0.046	0.039	11.8	6.3	16.0	6.9	0.8	1.9	0.6	2.4
S2 B8a	naive	0.009	0.004	0.049	0.039	47.3	4.9	47.9	5.1	1.3	2.0	1.2	2.2
	tamrfsits	0.016	0.016	0.043	0.036	13.6	6.6	14.5	6.7	5.5	1.6	5.6	2.1
S2 B11	naive	0.005	0.002	0.035	0.021	54.3	5.6	55.0	5.8	1.5	2.1	1.4	2.3
	tamrfsits	0.012	0.008	0.030	0.019	14.5	11.1	14.4	6.2	6.5	1.9	6.2	2.2
S2 B12	naive	0.005	0.002	0.029	0.019	52.9	6.2	53.8	6.4	1.5	2.2	1.4	2.4
	tamrfsits	0.010	0.007	0.026	0.017	15.8	10.6	15.4	6.3	6.3	2.1	6.0	2.3

Table 12: Full results table for the comparison between TAMRFSITS model and naive interpolation on the gap-filling task, where regular gaps of 30 days are masked from the input SITS and kept appart for validation. Only a selection of spectral bands is presented. Full results are available in appendix 5. ↓ (resp. ↑) indicates that the metric should be minimized (resp. maximized).

		RMSE↓		**BRISQUE↓**		**FR↑**	
Band	**Method**	mean	std	mean	std	mean	std
B5	dsen2	0.003	0.002	27.696	4.900	4.2	0.7
	tamrfsits	0.011	0.015	12.671	5.857	6.1	2.3
B6	dsen2	0.003	0.001	25.347	5.272	4.5	0.5
	tamrfsits	0.013	0.014	12.863	6.183	5.9	1.9
B7	dsen2	0.004	0.001	24.557	5.700	4.5	0.5
	tamrfsits	0.015	0.014	13.766	6.163	5.6	1.8
B8a	dsen2	0.004	0.001	24.906	5.477	4.5	0.4
	tamrfsits	0.015	0.014	13.940	6.501	5.6	1.7
B11	dsen2	0.002	0.001	33.227	5.204	3.9	0.6
	tamrfsits	0.012	0.010	14.788	10.035	6.5	1.8
B12	dsen2	0.002	0.001	32.182	4.992	4.0	0.7
	tamrfsits	0.010	0.009	15.783	9.712	6.4	2.0

Table 13: Full results table for the comparison between TAMRFSITS and DSen2 on the sharpening of Sentinel-2 20m bands, for a selection of bands. ↓ (resp. ↑) indicates that the metric should be minimized (resp. maximized).

		RMSE↓		**BRISQUE↓**		**FR↑**	
Band	**Method**	mean	std	mean	std	mean	std
B2	sen2like	0.028	0.040	30.9	19.5	0.1	3.6
	stair	0.014	0.008	21.2	8.2	0.9	2.9
	dh	0.028	0.013	54.7	9.3	-5.3	2.7
	dstfn	0.023	0.012	28.4	9.0	0.4	2.6
	tamrfsits (no hr)	0.015	0.007	49.8	6.8	-3.0	3.0
	tamrfsits (no lr)	0.013	0.008	25.1	10.9	0.5	2.9
	tamrfsits (full)	0.012	0.005	23.4	10.7	0.5	2.9
B3	sen2like	0.028	0.036	28.0	17.3	-0.4	3.3
	stair	0.016	0.008	17.0	5.3	0.7	2.6
	dh	0.027	0.015	50.7	7.6	-5.0	2.4
	dstfn	0.024	0.013	27.5	4.4	0.5	2.2
	tamrfsits (no hr)	0.017	0.008	47.4	5.5	-3.1	2.5
	tamrfsits (no lr)	0.015	0.009	19.3	6.2	0.5	2.5
	tamrfsits (full)	0.013	0.006	19.8	6.9	0.3	2.5
B4	sen2like	0.029	0.017	61.5	16.7	-3.4	2.6
	stair	0.018	0.010	18.9	6.4	0.5	2.1
	dh	0.027	0.015	49.0	6.3	-4.5	2.0
	dstfn	0.027	0.015	27.4	6.2	0.6	1.9
	tamrfsits (no hr)	0.020	0.009	45.9	5.3	-2.7	2.1
	tamrfsits (no lr)	0.017	0.010	19.4	7.3	0.4	2.1
	tamrfsits (full)	0.015	0.008	20.1	7.5	0.3	2.1
B8a	sen2like	0.048	0.033	58.4	10.6	-0.3	2.3
	stair	0.033	0.014	48.2	5.2	0.7	1.8
	dh	0.041	0.021	65.3	3.8	-1.9	1.5
	dstfn	0.047	0.033	23.3	5.7	3.3	2.1
	tamrfsits (no hr)	0.031	0.016	42.7	5.0	2.5	1.5
	tamrfsits (no lr)	0.029	0.015	14.7	7.0	5.1	1.4
	tamrfsits (full)	0.025	0.013	18.2	6.9	4.7	1.5
B11	sen2like	0.042	0.027	66.8	11.2	0.2	2.2
	stair	0.029	0.013	54.3	5.0	0.9	1.6
	dh	0.036	0.018	64.0	4.6	-1.3	1.6
	dstfn	0.041	0.027	27.0	8.0	4.1	2.0
	tamrfsits (no hr)	0.026	0.013	37.4	5.6	3.8	1.5
	tamrfsits (no lr)	0.026	0.013	15.1	5.9	5.8	1.5
	tamrfsits (full)	0.023	0.011	17.7	6.3	5.4	1.4
B12	sen2like	0.036	0.020	65.8	12.5	0.2	2.0
	stair	0.025	0.013	52.5	5.1	0.8	1.6
	dh	0.029	0.016	64.1	5.0	-1.4	1.5
	dstfn	0.035	0.020	29.2	7.6	4.0	1.7
	tamrfsits (no hr)	0.021	0.011	37.9	5.3	3.5	1.5
	tamrfsits (no lr)	0.022	0.012	15.5	5.8	5.6	1.5
	tamrfsits (full)	0.019	0.010	18.1	6.2	5.3	1.4

Table 14: Full results table of the comparison of the different methods on the spatio-temporal fusion task. Only a subset of bands is presented. Full results are available in appendix 5. * if usually higher FR means higher spatial resolution details, in the case of bands that are 10 meter native resolution, the best FR is the closest to zero in absolute value. Otherwise, ↓ (resp. ↑) indicates that the metric should be minimized (resp. maximized).