Temporal ViT-U-Net Tandem Model: Enhancing Multi-Sensor Land Cover Classification Through Transformer-Based Utilization of Satellite Image Time Series

Mohammadreza Heidarianbaei[1] , Hubert Kanyamahanga[1] , Mareike Dorozynski[1]

[1] Institute of Photogrammetry and GeoInformation, Leibniz University Hannover, Germany
mohammadreza.heidarianbaei@stud.uni-hannover.de
(kanyamahanga, dorozynski)@ipi.uni-hannover.de

Keywords: Semantic segmentation, Land cover classification , Multi-Sensor remote sensing , Vision transformer , Satellite image time series

Abstract

Semantic segmentation is essential in the field of remote sensing because it is used for various applications such as environmental monitoring and land cover classification. Recent advancements aim to collectively classify data from diverse sensors and epochs to improve predictive accuracy. With the availability of vast Satellite Image Time Series (SITS) data, supervised deep learning methods, such as Transformer models, become viable options. This paper introduces the Temporal Vision Transformer(ViT), designed to extract features from SITS. These features, capturing the temporal patterns of land cover classes, are integrated with features derived from aerial imagery to improve land cover classification. Drawing inspiration from the success of transformers in Natural language processing (NLP), Temporal ViT concurrently extracts spatial and temporal information from SITS data using tailored positional encoding strategies. The proposed approach fosters comprehensive feature learning across both domains, facilitating seamless integration of encoded data from SITS into aerial images. Furthermore, a training strategy is proposed that supports the Temporal ViT to focus on classes with a changing appearance over the year. Extensive experiments carried out in this work indicate the enhanced classification performance of Temporal ViT compared to existing state-of-the-art techniques for multi-modal land cover classification. Our model achieves a 3.8% increase in the mean IoU compared to the network solely relying on aerial images.

1. Introduction

Semantic segmentation is a task in photogrammetry, remote sensing, and computer vision in which a class label is assigned to each pixel in the image. In the field of remote sensing, semantic segmentation has a long history and has always been key to extracting detailed information from satellite or aerial imagery for various applications such as environmental monitoring and land cover analysis (Blaschke et al., 2000; Yuan and Sarma, 2010; Yang et al., 2016). Semantic segmentation based on a single data source such as aerial or satellite images from a single point in time has been extensively investigated (Marmanis et al., 2016; Favorskaya and Zotin, 2021; Niu et al., 2021) Many works have shown that the results of the classification can be improved by combining data from multiple sensors (Benedetti et al., 2018; Bergamasco et al., 2023; Garioud et al., 2023) . Over the past few years, there has been a growing interest in jointly classifying data from different sensors, and time steps (Benedetti et al., 2018; Bergamasco et al., 2023; Garioud et al., 2023; Yan et al., 2023). The heterogeneous nature of remote sensing data, which includes variations in spectral, spatial, and temporal resolution, comes along with both strengths and limitations for each type of data. Aerial imagery, with its remarkable spatial resolution, faces challenges in temporal frequency due to the high costs of acquisition and the lack of consistently available systems for capturing these images. This limitation obstructs the thorough representation of temporal object characteristics in remote sensing data, particularly affecting the accurate distinction of classes such as vegetation, which undergo varying appearances over time. Conversely, SITS data provide high temporal frequency and the capacity to capture temporal changes. This enables models to learn temporal patterns from such data. However, it does suffer from lower spatial resolution, which compromises the precise distinction of highly detailed objects. To benefit from the respective modality-specific (in this paper, by different modalities, we specifically mean aerial and satellite imagery) advantages, both satellite and aerial images can be combined to improve the classification of land cover.

The main objective of this paper is to present a method that combines SITS and mono-temporal aerial images to obtain a land cover map at the spatial resolution of the aerial image while exploiting temporal information contained in SITS. Given the remarkable achievements of deep learning techniques (Yuan et al., 2020), current approaches dedicated to the semantic segmentation of multi-sensor remote sensing data are predominantly based on such deep learning architectures (Ienco et al., 2019; Garioud et al., 2023; Yan et al., 2023). The recent success of transformers (Vaswani et al., 2017) in Natural Language Processing has been extended to computer vision tasks. Vision Transformer (ViT)(Dosovitskiy et al., 2021) adapts transformer-based architectures originally used for NLP to process images by treating them as sequences of patches rather than pixels, pioneering their application in image classification. By partitioning an image into patches that are considered as tokens, ViT facilitates the adaptation of transformer models to vision-related tasks. However, very few research studies have used self-attention approaches to jointly integrate SITS and aerial images in classifying land cover (Garioud et al., 2023). The study of (Garioud et al., 2023) incorporated a self-attention module to process SITS only in the temporal dimension. The results show improvements by integrating temporal information compared to exclusively utilizing aerial images for land cover classification. Leveraging the core concept of ViT, ViT variants have demonstrated effectiveness in handling SITS data, particularly for the task of semantic segmentation. (Tarasiou

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-3-2024-169-2024 | © Author(s) 2024. CC BY 4.0 License.

et al., 2023) introduced Temporal-Spatial Vision Transformer (TSViT) for SITS network for handling SITS data, and they demonstrated that the order of factorization; i.e., extracting temporal then spatial information from SITS data is more important for downstream tasks such as semantic segmentation. This network architecture showcases the promising utilization of transformer blocks in synthesizing SITS data, thus offering an avenue to enrich spatial details extracted from aerial imagery by incorporating temporal information extracted from SITS. Previous approaches for integrating self-attention into SITS data (Tarasiou et al., 2023; Voelsen et al., 2023) have often followed a two-step approach. Such approaches generally involve encoding images separately, first in the spatial domain, then in the temporal domain, or vice versa. While processing spatial and temporal information concurrently enables a model to capture intricate interactions and obtain comprehensive representations of data spanning both domains, the sequential nature of existing approaches introduces challenges in directly computing attention weights across both domains, potentially impeding effective encoding. The present work aims to overcome this limitation.

Motivated by the improvements demonstrated in prior research employing transformers for extracting temporal features from SITS data, we introduce our Temporal ViT network. Drawing inspiration from self-supervised pre-training (Cong et al., 2022), the positional encoding is adapted in both temporal and spatial dimensions. This enables collective tokenization of the entire set of time series input, which is subsequently processed by a ViT encoder. Such simultaneous processing of spatial and temporal data promotes comprehensive global receptive fields in both domains. Consequently, our approach facilitates feature learning across both temporal and spatial dimensions of SITS. The extracted features from SITS data are seamlessly combined with features learned from aerial images, enhancing the feature maps of the latter with encoded information from SITS to improve the final land cover classification. The scientific contributions of this work can be summarized as follows:

- We propose a Temporal ViT encoder for SITS so that we can combine spatial and temporal features from SITS with aerial imagery to improve the results of land cover classification. At the core of the Temporal ViT are a spatio-temporal positional encoding and an additional learnable classification token, from which we derive the pixel-wise label map.
- We investigate various positional encoding schemes and show that integrating both the spatial and temporal position of each patch in the image of the SITS data can improve the performance of the classifier in correctly predicting land cover classes.
- Furthermore, a training strategy is proposed that guides the Temporal ViT to focus on classes with a changing appearance in SITS, while simultaneously detailed geometrical information about all classes is forced to be extracted from aerial images, aiming at high-quality multi-modal land cover prediction.

2. Related Work

In this section, we review related work that focuses on the integration of multi-sensor and multi-temporal data for classification. We start by looking at existing approaches that use conventional machine learning models, then we introduce models based on convolutional neural networks (CNNs; (LeCun et al., 1989; Krizhevsky et al., 2012)) and how they are used to jointly combine data from multiple sensors. Finally, we discuss the few existing approaches that utilize attention-based modules to jointly fuse aerial and satellite image time series data.

2.1 Classical machine learning methods

Machine learning methods have been used to classify multi-modal data. For instance, Campos-Taberner et al. (2019), stacked multi-temporal Sentinel-2 and Sentinel-1 images, which were then utilized to train diverse traditional machine learning classifiers, including decision tree ensembles and Support Vector Machines. Moreover, to enhance crop type prediction, a discriminative linear chain Conditional Random Field is employed to model temporal dependencies (Giordano et al., 2018). However, such traditional models rely on hand-crafted features extracted from the input images, which is why they have mostly been supplanted by deep learning architectures that can learn to extract relevant features through convolutional layers, leading to better performance for the task of classification (Rußwurm and Körner, 2018; Turkoglu et al., 2021).

2.2 Deep learning methods

Fully Convolutional Networks have frequently been employed in remote sensing applications for the task of semantic segmentation (Long et al., 2015; Marmanis et al., 2016; Ma et al., 2021; Voelsen et al., 2022). Those networks utilize convolutional kernels with learnable parameters that are shifted across the input data to extract discriminative features. In the realm of multi-modal data, two different principles have been used to fuse data from multiple sources. On the one hand, features from different sources can be combined and provided as a joint input to a network for further processing. For instance, in (Adrian et al., 2021), Principal Component Analysis is employed in Sentinel-2 and Sentinel-1 images to convert the original data into a new set of features. The resultant features are then used as inputs to a SegNet (Badrinarayanan et al., 2017) to generate pixel-wise label maps at the geometrical resolution of Sentinel-2 imagery. This approach is recommended when the inputs from both data sources have the same spatial resolution. On the other hand, other works adopted a two-branch architecture where each data modality is separately processed by a dedicated CNN-based network. In (Benedetti et al., 2018), a Gated Recurrent Unit network is used to process Sentinel-2 time series data and a 2D-CNN network to extract spatial features from SPOT-6 mono-temporal images. Features computed separately from both branches are then concatenated to predict land cover at the spatial resolution of Sentinel-2 imagery.

Very few CNN-based approaches combine aerial images and SITS data for pixel-wise classification. In (Bergamasco et al., 2023), a 3D-CNN network is utilized to extract spatio-temporal features from Sentinel-2 SITS data which are combined with features learned from aerial images processed by a 2D-ResNet (He et al., 2016). The extracted features from each modality are then combined and used as input to the decoder for the final classification (Bergamasco et al., 2023). The results show that combining aerial images and SITS data improves the classification results compared to uni-modal classification. However, the method shows limitations in distinguishing classes that are semantically similar.

2.3 Attention-based models for SITS

Attention mechanisms, initially developed in the field of natural language processing (Vaswani et al., 2017), have been effectively adopted in computer vision with the emergence of ViT. By substituting convolutional layers with self-attention modules, ViTs facilitate the comprehensive modeling of interdependencies among image patches on a global scale. Inspired by ViT, the original ViT network was adapted to handle SITS data (Tarasiou et al., 2023). Each image in the SITS is divided into non-overlapping patches and multiple ViT blocks are executed in parallel for each timestep, where attentions are computed between all timesteps of the corresponding patches at the same spatial location in the image. Afterward, the outputs are reshaped and the attentions are computed between all patches of the same timestep to learn spatio-temporal characteristics of SITS data. Following a different method, in (Voelsen et al., 2023), a spatio-temporal transformer block (ST-TB) is introduced in the standard Swin Transformer (Liu et al., 2021a), initially designed for mono-temporal images, to adapt it to SITS data. ST-TB is employed in conjunction with the standard Swin-Transformer blocks (STB), where the parallel STBs encode individual timesteps. Subsequently, the outputs of all time steps are fused and encoded by the ST-TB. The two approaches mentioned above fall short of establishing direct global receptive fields in time and space, which prevents the use of interactions and the acquisition of representations covering spatial and temporal domains. Although factorizing SITS can be computationally efficient, this technique is more appropriate for data types like videos (Arnab et al., 2021), which typically have higher spatial and temporal resolutions. Tokenization and encoding of all temporal epochs of data with such high spatial and temporal resolutions collectively would result in a large number of tokens, potentially making it computationally expensive due to the quadratic complexity of the attention mechanism. However, this challenge is less significant with satellite images since they usually have lower spatial and temporal resolutions than videos.

Distinguished from two-step encoding methodologies, Gao et al. (2022) introduces a paradigm facilitating space-time attention through Cuboid Attention. By extracting tokens from spatio-temporal input cuboids, this method encodes data. Nevertheless, it is important to highlight that this methodology still lacks direct interaction among patches from diverse spatial and temporal positions, and further it does not address semantic segmentation. Masked autoencoders (MAEs) (He et al., 2022) have emerged as a powerful paradigm in self-supervised learning, capable of learning rich representations by reconstructing masked input data. Building upon this foundation,SatMAE (Cong et al., 2022) employs MAEs to handle temporal and multi-spectral input data effectively. SatMAE incorporates a positional encoding for the spatio-temporal or spatio-spectral dimensions, enabling attention in the respective two domains. However, it’s essential to note that SatMAE’s primary focus remains on self-supervised learning and is not suitable for tasks requiring the generation of pixel-wise label maps from SITS. Irrespective of the approach employed in leveraging ViT variants for processing SITS data, there is a need to design a method for encoding SITS with ViT and merging this feature map with single-time aerial images to enhance the network’s capacity to attain land cover classification prediction.

A work that is relatively close to our approach to integrating SITS data using an attention-based approach for multi-modal data is (Garioud et al., 2023), where a two-branch architecture based on U-Net is used to fuse SITS and aerial images. One U-Net network with lightweight temporal attention (L-TAE; (Garnot and Landrieu, 2020)) is used for the SITS branch and a standard U-Net is used for aerial images. A fusion module is inserted to merge features resulting from both branches. The module first refines information from SITS data before being injected through skip connections in the network processing aerial images to finally produce multi-modal land cover predictions. Furthermore, (Kanyamahanga and Rottensteiner, 2024) examine the effectiveness of the Swin Transformer (Liu et al., 2021b) and the method proposed by (Tarasiou et al., 2023) as SITS encoders within the framework of (Garioud et al., 2023). However, their approach lacks simultaneous encoding in the spatio-temporal dimensions.

To the best of our knowledge, none of the existing methods have examined the utilization of transformer-based models to joint extraction of spatio-temporal features in the multi-modal land cover classification task. In this paper, we extend the approach in (Garioud et al., 2023) by introducing a transformer model, Temporal ViT, for processing SITS. Unlike previous works, where spatial and temporal features are extracted one after another, our approach extends ViT, enabling direct and simultaneous learning of spatio-temporal representations. Moreover, we propose a training strategy for the Temporal ViT that aims at focusing on classes with a changing appearance to extract the most information through SITS for multi-modal land cover classification.

3. Methodology

The main goal of our method is to use SITS data and an aerial image to predict one land cover map at the spatial resolution of the aerial image while exploiting the temporal information contained in the SITS. The key idea of our approach is to use a spatio-temporal positional encoding strategy to effectively leverage the information contained in SITS data. The proposed architecture is based on the one proposed in (Garioud et al., 2023), where our novelty is the design of a fully transformer-based network for encoding SITS data. Thus, we aim to extract representative features of objects with a changing appearance in a better way.

The proposed method, as depicted in Figure 1, consists of two components: one for processing aerial images, denoted as $Aerial\ branch$, and one for processing SITS denoted as $Satellite\ branch$. Those two branches are connected by a fusion module adopted from (Garioud et al., 2023) to fuse features learned from both branches which are afterward used to produce a pixel-wise label map at the spatial resolution of the aerial image. The $Aerial\ branch$ is identical to the one in (Garioud et al., 2023). In Section 3.1, a brief description of the Temporal ViT, which is the novelty in this paper is provided. More precisely, our first contribution is a spatio-temporal positional encoding in the context of supervised learning, which is described in Section 3.1.1. The end-to-end encoder-decoder structure of our network is described in Section 3.1.2, where the SITS decoder is our second contribution to the model architecture. The training procedure used for our approach is described in Section 3.2.

3.1 Temporal ViT for SITS data

In this part, we focus on the $Satellite\ branch$ also presented in Figure 1. The input to the Temporal ViT is defined as $X_{sat} \in$

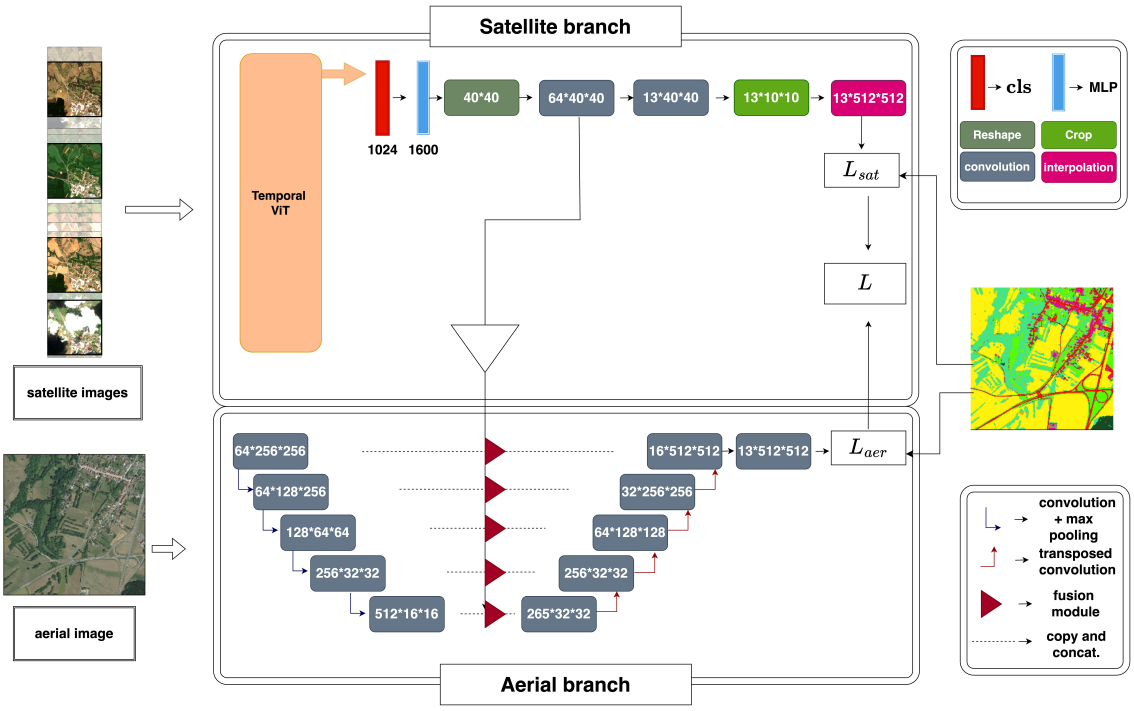

Figure 1. Network architecture for the joint classification of aerial and of SITS data, adapted from Garioud et al. (2023) . The model architecture consists of two main branches: the satellite branch and the aerial branch. In the satellite branch, the proposed Temporal ViT is utilized to encode SITS. Meanwhile, the aerial branch processes aerial image. The encoded SITS data is then fused with the output of the U-Net encoder through the skip connections, facilitated by a fusion module adopted from Garioud et al. (2023).

$\mathbb{R}^{T\times C_{sat}\times H_{sat}\times W_{sat}}$, i.e. T satellite image time steps X of size $H_{sat} \times W_{sat}$ with C_{sat} channels. Similar to ViT, each satellite image $X \in X_{sat}$ is partitioned into N non-overlapping patches of size $N \times (B^2 \times C_{sat})$ as shown in Figure 3. Here N represents the total number of patches in each image and B denotes the number of pixels in width and height of each patch.

The resulting sequence of flattened time series images is then concatenated constituting an input vector with dimensions $V \times (B^2 \times C_{sat})$, where $V = T \cdot N$ denotes the total number of patches in the time series. Linear transformations are applied to these flattened patches to generate the patch embeddings $e_v \in \mathbb{R}^d$ with $v = 1, ..., V$, where d denotes the embedding dimension. Recognizing the permutation invariance of transformer models, it becomes essential to incorporate positional information into the input, ensuring the network's awareness of spatial and temporal relationships. Unlike standard ViT, originally designed to handle one image, where the positional encoding encodes the position of each patch within the image and thus provides the model with the required information with respect to the order of token, dealing with sequences of time series requires a different tokenization approach. To achieve this, a spatial-temporal positional encoding is adapted from (Cong et al., 2022). In the rest of this paper, we refer to PE_{S-T} as the spatial-temporal positional encoding, as described in Section 3.1.1. Moreover, a learnable token, denoted as $\mathbf{cls} \in \mathbb{R}^d$, is concatenated to the input patch embeddings to aggregate all feature maps from the patches and facilitate the generation of the pixel-wise label map at the end of the network. The input sequence undergoes processing by the transformer encoder, adhering to the method outlined in (Vaswani et al., 2017). The generation of the pixel-wise label map is deduced from the output class token, originating from the transformed encoder, as elaborated in Section 3.1.2.

3.1.1 Spatio-Temporal Encoding

The goal of the spatio-temporal encoding PE_{S-T} is to encapsulate information about the spatial and temporal positions of each patch in each image in a SITS, which are concatenated together as presented in Figure 2. The mathematical formulation is given as follows:

$$PE_{S-T}(t, s) = \begin{bmatrix} PE_{temp}(t) \\ PE_{spatial}(s) \end{bmatrix} \tag{1}$$

In Equation 1, PE_{temp} denotes the temporal positional encoding (see Equation 2), and the spatial positional encoding $PE_{spatial}$ follows the original ViT positional encoding.

With s in $PE_{spatial}$ denotes the spatial position of the patch within the image. Where in PE_{temp}, the t is defined by the elapsed number of days since the first image acquisition, where the initial satellite image serves as the temporal reference point, i.e. $day = 0$. This specification guarantees the model's access to accurate and informative prior knowledge about the temporal position of each patch in the sequence, where the encoding is defined as follows:

$$PE_{temp}(day, i) = \begin{cases} \sin\left(\frac{day}{10000^{2i/d}}\right), & \text{if } i \text{ is even} \\ \cos\left(\frac{day}{10000^{2i/d}}\right), & \text{if } i \text{ is odd} \end{cases} \tag{2}$$

Here, d is the dimensionality of the input embeddings, i is the dimension index with $i \in \{0, ..., \frac{d}{3} - 1\}$. The length of the temporal positional encoding is defined to be one-third of the spatio-temporal encoding, as proposed in (Cong et al., 2022), and $10,000$ is a scaling factor.

As mentioned above, the spatio-temporal positional encoding PE_{S-T} is added to the patch embeddings e_v, such that the input to the transformer encoder can be represented by

https://doi.org/10.5194/isprs-annals-X-3-2024-169-2024

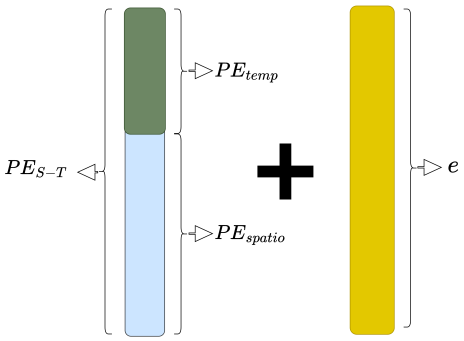

Figure 2. Spatio-temporal positional encoding PE_{S-T} is constructed by concatenating spatial positional encoding $PE_{spatial}$ (blue) and temporal positional encoding PE_{temp} (green) , as illustrated in the figure. Here, e represents the embedded patch.

$$E = \{e_v + PE_{S-T}(t, s)\}_{t=1,\ldots,T;s=1,\ldots,N} \quad (3)$$

where E represents a set of all input tokens. the Thus, the model gains the ability to discern not only spatial but also temporal relationships between images captured at different acquisition times.

3.1.2 Encoder - Decoder To grasp the global context across SITS and abstract both spatial and temporal information, similar to ViT, we add a (**cls**) token to the sequence of patch embedding (**E**), the result of which is provided to the encoder where the attention mechanism weighs the importance of each token in the input sequence to every other token in the input sequence enabling the encoder to capture spatial-temporal patterns. The encoder produces a $(1 + V) \times d$-dimensional output, which encodes SITS information. Only the class token, $\mathbf{cls} \in \mathbb{R}^d$, is preserved from the output sequence as a singular representation of the learned feature map of the input sequence. Further, these features have to be redistributed in the spatial domain, such that a $H_{sat} \times W_{sat}$ dimensional feature map is obtained to be integrated with aerial imagery. To do so, this class token is presented to a multi-layer perceptron (MLP), realizing a transformation $f : \mathbb{R}^d \rightarrow \mathbb{R}^{H_{\text{sat}} \cdot W_{\text{sat}}}$, such that a new vector suitable for subsequent operations is obtained. Following this, the resulting vector undergoes a reshape operation to generate a two-dimensional feature map of dimensions $H_{\text{sat}} \times W_{\text{sat}}$, which fulfills the desired spatial requirements. Subsequently, the feature map obtained is further refined through convolutional layers, resulting in C feature map dimensions of the U-Net model. Finally, the outputs of the $Satellite\ branch$ are fused with the aerial features learned by the U-Net encoder, processing aerial images. Both the fusion of SITS feature maps and the extraction of aerial features by a U-Net are identical to (Garioud et al., 2023).

3.2 Network Training

To train our network, a loss function L comprises a weighted sum of two losses, i.e. one for the $Aerial\ branch$ (L_{aer}) and one for $Satellite\ branch$ (L_{sat}). The total loss is minimized using mini-batch Stochastic Gradient Decent (Ruder, 2016). Both loss functions, L_{aer} and L_{sat}, are based on the categorical Cross Entropy loss, where L_{aer} is selected to be the standard variant of that loss. In contrast, weights are considered in L_{sat}. The loss terms in L_{sat} belonging to pixels i of a certain class k are equally weighted, except for loss terms belonging to pixels

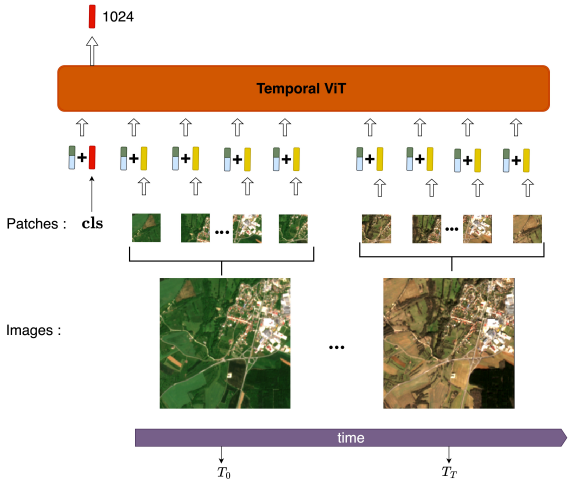

Figure 3. Temporal ViT processes a sequence of SITS images, denoted as t_0 to t_T, as input. Each image is partitioned into non-overlapping patches, which are then linearly projected as depicted by the yellow blocks in Figure. Spatio-temporal positional embeddings are incorporated into these projected patches. A cls token is is added to the beginning of the sequence before feeding it into the Transformer encoder. This token, representing the final output of the Transformer, has a dimensionality of 1024.

of the set of classes C_{static}, where no changes in their appearance are expected over time; the loss weight of such pixels is set to zero, i.e. they do not contribute to the weight update of the SITS encoder, encouraging to focus on dynamics in appearance. Thus, L_{sat} becomes

$$\mathcal{L}_{sat} = -\sum_{i=1}^{P}\sum_{k=1}^{K} y_{ik} \log(p_{ik}) w_{ik} \quad (4)$$

where w_{ik} is weight belonging to pixels i for a class k, and $w_{ik} = 0$ in case the i^{th} sample belongs to a class $k \in C_{static}$ and $w_{ik} = 1$ in all other cases. M is the total number of pixels in the mini-batch, and K represents the total number of classes. The binary indicator variable y_{ik} denotes whether the target class label of a pixel i is k ($y_{ik} = 1$) or not ($y_{ik} = 0$). p_{ik} is the softmax output for pixel i to belong to class k.

As the ground truth labels are only provided for aerial images, covering a smaller area compared to SITS, the predictions of the SITS branch are first cropped to the small area and then upsampled to the spatial extent of the aerial image through bilinear interpolation before computing the loss $\mathcal{L}_{\text{branch}}$ for Satellite branch. For the aerial branch, the parameters of the network are initialized starting from the weights pre-trained on the ImageNet dataset (Russakovsky et al., 2015), while for the SITS branch, the parameters are randomly initialized.

During inference, only the aerial branch predictions are considered as the final output. L_{sat} softmax score serve as an auxiliary loss function only during the training phase.

4. Experiments & Results

4.1 Dataset

We use the French Land cover from Aerospace ImageRy (FLAIR) dataset (Garioud et al., 2023), consisting of mono-

temporal aerial imagery acquired between April 2018 and November 2021 and Sentinel-2 time series images over one year tailored for semantic segmentation tasks. The dataset contains 77,762 image patches, each with an aerial image, a SITS, and a ground truth label map, where the patches are distributed across 50 regions in France. For aerial images, patches of 512 x 512 pixels with a ground sampling distance (GSD) of 0.2m are used. Each aerial image contains four channels (red, green, blue, and near-infrared) and a digital surface model. For SITS, an image patch of 40 x 40 pixels is considered at a GSD of 10m. Bands originally captured at a ground sampling distance of 20m are resampled to 10m using bilinear interpolation. The time series covers the entire year during which the corresponding aerial image was acquired and comprises between 20 and 110 images. Pixel-wise annotations are provided for each patch at the spatial resolution of the aerial image. These annotations distinguish between 13 classes, namely: *building*, *pervious surface*, *impervious surface*, *bare soil*, *water*, *coniferous*, *deciduous*, *brushwood*, *vineyard*, *herbaceous vegetation*, *agricultural land*, *plowed land* and *other*.

4.2 Experimental Setup

Following the splitting protocol established in (Garioud et al., 2022), the dataset is divided into three subsets: training, validation, and test set, containing 48,812, 12,900, and 16,050 patches, respectively, summing up to a total of 77,762 patches. We apply various data augmentation strategies to the training dataset, namely rotations by 90°, 180°, 270°, horizontal and vertical flipping, as well as color augmentation which randomly alters the brightness and contrast of images which results in a wider range of input variations. In our training procedure, the batch size used for each epoch is fixed to 4, and the learning rate is set to 0.01. Furthermore, we utilize a learning rate decay schedule, reducing the learning rate by a factor of 0.1 every 10 epochs. However, to further prevent over-fitting, early stopping is employed; training is stopped if there is no decrease in the validation set loss over 30 consecutive epochs. Regarding the hyper-parameters of ViT, we opt for patch of the size $B = 16$ pixels, a depth of 24 (number of transformer blocks), and employed 16 self-attention heads. All these hyper-parameters were selected based on their performance on the validation dataset. The hyper-parameters of the aerial branch are identical to those in (Garioud et al., 2023). Our method was implemented using PyTorch Lightning framework (Falcon and The PyTorch Lightning team, 2019) and the training is carried out on 2 Nvidia A100 GPUs.

We conduct three sets of experiments. In the first set of experiments, we use the U-Net (Garioud et al., 2022), to predict land cover only based on the aerial images; its results are compared against those of the other methods, which use SITS data and aerial images, to assess the impact of the SITS on the classification. To evaluate the influence of Temporal ViT, we also consider the second set of experiments based on the U-TAE and U-Net architectures (U-T&T) as detailed by (Garioud et al., 2023), which incorporates both aerial images and SITS, as our baseline. The last set of experiments, referred to as Temporal ViT are described in section 3.1. Moreover, we conducted a series of experiments employing diverse positional encoding schemes to assess the impact of spatio-temporal positional encoding on our proposed Temporal ViT network. Here, the class *plowed land* is considered in the set C_{static} for training the Temporal ViT; all other classes are either considered to underlay visual changes over time in SITS or are not considered in that, because it is assumed that they cannot be recognized in utilized satellite images, such as *building*.

To provide a more accurate representation of the model's performance, each experiment is repeated three times, each time starting from a different random initialization of the weights and using random shuffling for batches, to assess the impact of these random components on the classification results.

To assess the performance of the conducted experiments, the classification results computed on the test image patches are compared to the ground truth labels, and the intersection over union (IoU_c) of each class is reported:

$$IoU_c = \frac{TP_c}{TP_c + FP_c + FN_c} \tag{5}$$

TP_c, FP_c, and FN_c denote the number of pixels that are true positives, false positives, and false negatives, respectively, for a class c. We also report the mean intersection over union ($mIoU$), which is computed by taking the mean of the (IoU_c) values of all classes excluding the class ($other$) despite its contribution to the loss function.

4.3 Results & Discussion

In this section, we present the results achieved by different networks described in the previous section 4.2.

4.3.1 Impact of multi-temporal satellite imagery Table 1 shows the IoU_c and respective $mIoU$ achieved in the first three sets of experiments described above, demonstrating the impact of SITS data on classification and in particular, of the proposed Temporal ViT. The numbers show that the use of SITS data as an additional source of information leads to an increase in overall performance compared to U-Net; when using only aerial imagery with U-Net, a $mIoU$ score of 54.5% is achieved. The U-T&T model, incorporating SITS data, yielded improved $mIoU$ of 56.0%, indicating a statistically significant enhancement over the baseline U-Net model. Additionally, our Temporal ViT model achieved a slightly better $mIoU$ score of 58.1%, demonstrating the advantage of the proposed SITS encoder. These results are further supported by hypothesis testing (t-test), which confirmed the statistically significant improvement in performance at a 95% confidence level. The improvement underscores the significance of learning spatio-temporal features, made feasible by employing spatio-temporal positional encoding that takes into account the acquisition date of the SITS data, as well as by the utilized training strategy. Notably, while U-Net comprises 27.5 million parameters and U-T&T 33.5 million, our Temporal ViT model features a substantially larger parameter count of 330 million, indicating a significant increase in model parameters, which contributes to its superior performance.

By looking at the class-specific metrics, It can be seen that the U-Net model outperforms other models with a large margin of (6.1%) on only one class *herbaceous vegetation*. *Herbaceous vegetation*, found in locations like gardens, public parks, and recreational fields utilized for sports, due to the heterogeneous type of this class, doesn't follow distinct temporal and spatial patterns in satellite images. Consequently, the feature map derived from the temporal branch for this specific class may contain misleading data.

On the other hand, Temporal ViT yielded better performance, notably for classes that evolve with temporal variations over

Class	U-Net	U-T&T	Temporal ViT (Ours)
building	81.5 ± 0.2	80.9 ± 0.4	**82.0** ± 1.0
pervious surface	49.5 ± 0.1	49.1 ± 1.1	**52.7** ± 1.0
impervious surface	**72.5** ± 0.2	71.0 ± 0.2	72.0 ± 1.0
bare soil	41.3 ± 3.0	39.8 ± 5.0	**49.8** ± 5.0
water	83.4 ± 2.0	83.0 ± 1.9	**85.2** ± 0.7
coniferous	35.8 ± 7.0	58.2 ± 3.5	**63.6** ± 2.0
deciduous	67.2 ± 2.0	70.2 ± 1.9	**71.1** ± 1.0
brushwood	23.4 ± 0.7	24.2 ± 2.5	**24.3** ± 3.0
vineyard	62.0 ± 0.7	**63.1** ± 2.9	62.9 ± 2.0
herbaceous vegetation	**48.9** ± 0.4	41.4 ± 5.2	42.8 ± 3.0
agricultural land	50.7 ± 2.0	52.9 ± 0.9	**54.5** ± 2.0
plowed land	**37.8** ± 2.0	37.0 ± 2.0	36.8 ± 2.0
mIoU [%]	54.5 ± 1.0	56.0 ± 0.9	**58.1** ± 1.0

Table 1. Mean Class-wise IoU values [%] and additionally, the corresponding standard deviations obtained from repeated training sessions. On the test set of the FLAIR #2 dataset produced by different methods. The models compared here are U-Net model trained solely on aerial images, U-T&T, the baseline from (Garioud et al., 2023), and Temporal ViT, our model which incorporates both aerial and SITS data.

Model	Input	PE	mIoU [%]
U-Net	aerial	-	54.5 ± 1.0
U-T&T	aerial + sat	Temporal	56.0 ± 0.9
Temporal ViT	aerial + sat	Spatial	50.4 ± 1.1
Temporal ViT	aerial + sat	Learnable	54.2 ± 0.8
Temporal ViT (Ours)	aerial + sat	Spatio-Temporal	**58.1** ± 1.0

Table 2. Semantic segmentation results achieved by using U-Net (Garioud et al., 2022) and U-T&T (Garioud et al., 2024) as baseline models, alongside our Temporal ViT model with different positional encoding schemes. PE refers to the type of positional encoding. The mIoU and standard deviation are derived from repeated training sessions.

the year, such as *bare soil*, *coniferous*, and *agriculture land*, where accuracy is significantly improved. These results underline the importance of temporal information contained in these classes, which is captured by the Temporal ViT and U-T&T, where Temporal ViT significantly outperformed U-T&T.

4.3.2 Impact of Spatio-Temporal Positional Encoding We investigate the influence of spatio-temporal positional encoding, by running multiple experiments with different types of positional encoding, the results of which are presented in Table 2. Initially, we introduced $PE_{spatial}$ where only the spatial position of each patch in the image is considered, and a $mIoU$ of 50.4% is achieved. Following this, we implemented a learnable positional encoding mechanism (Gehring et al., 2017), treating positional encoding vectors as model parameters. These parameters were then updated during training, alongside the remainder of the model parameters. This adaptive approach resulted in an improvement of 3.8%, compared to a fixed spatial positional encoding. Finally, we applied PE_{S-T} encoding, which takes into account the spatial and temporal position of each patch in the SITS, and achieved a significance improvement of 7.7%, i.e., 58.1% $mIoU$. As can be expected, incorporating the relative or absolute spatio-temporal position of each image patch helps the classifier to gain a better understanding of the context in which pixels evolve, which in return, can be used to separate pixels that may have similar characteristics, even though they belong to two different classes.

5. Conclusions & Outlook

In this work, we proposed a method named Temporal ViT, capable of simultaneously learning feature representations in spatial and temporal domains to integrate aerial images and SITS data for multi-modal land cover classification. The experimental findings indicate an enhancement in performance through the incorporation of SITS data, yielding a 3.8% increase in $mIoU$ compared to U-Net networks relying solely on aerial images. While the U-T&T network has previously demonstrated the efficacy of SITS data integration, our work showcases that leveraging Temporal ViT for encoding SITS further boosts land cover classification accuracy by an additional 2.1% over the U-T&T model.

Nevertheless, the quadratic complexity inherent in our attention mechanism imposes constraints on the spatial and temporal resolution of input time series images, potentially leading to hardware bottlenecks and extensive GPU usage. Future research endeavors could focus on optimizing our method for satellite imagery with smaller GSD and higher temporal resolution. Furthermore, incorporating self-supervised techniques for pre-training, which has demonstrated improvements in ViT (Wang et al., 2023), could be explored. In addition, extending our transformer-based approach to aerial imagery and using it in combination with SITS data constitutes an intriguing avenue for future research.

References

Adrian, J., Sagan, V., Maimaitijiang, M., 2021. Sentinel SAR-optical fusion for crop type mapping using deep learning and Google Earth Engine. *ISPRS Journal of Photogrammetry and Remote Sensing*, 175, 215–235.

Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C., 2021. Vivit: A video vision transformer. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 6836–6846.

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE TPAMI*, 39(12), 2481–2495.

Benedetti, P., Ienco, D., Gaetano, R., Ose, K., Pensa, R. G., Dupuy, S., 2018. M3Fusion: A Deep Learning Architecture for Multi-{Scale/Modal/Temporal}

satellite data fusion. *ArXiv*, abs/1803.01945. https://api.semanticscholar.org/CorpusID:3744872.

Bergamasco, L., Bovolo, F., Bruzzone, L., 2023. A Dual-Branch Deep Learning Architecture for Multisensor and Multitemporal Remote Sensing Semantic Segmentation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16, 2147-2162.

Blaschke, T., Lang, S., Lorup, E., Strobl, J., Zeil, P., 2000. Object-oriented image processing in an integrated GIS/remote sensing environment and perspectives for environmental applications. *Environmental information for planning, politics and the public*, 2(1995), 555–570.

Campos-Taberner, M., García-Haro, F. J., Martínez, B., Sánchez-Ruíz, S., Gilabert, M. A., 2019. A Copernicus Sentinel-1 and Sentinel-2 Classification Framework for the 2020+ European Common Agricultural Policy: A Case Study in València (Spain). *Agronomy*, 9(9), 556.

Cong, Y., Khanna, S., Meng, C., Liu, P., Rozi, E., He, Y., Burke, M., Lobell, D. B., Ermon, S., 2022. SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery. A. H. Oh, A. Agarwal, D. Belgrave, K. Cho (eds), *Advances in Neural Information Processing Systems*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An image is worth 16x16 words: Transformers for image recognition at scale.

Falcon, W., The PyTorch Lightning team, 2019. PyTorch Lightning.

Favorskaya, M. N., Zotin, A. G., 2021. Semantic segmentation of multispectral satellite images for land use analysis based on embedded information. *Procedia Computer Science*, 192, 1504–1513.

Gao, Z., Shi, X., Wang, H., Zhu, Y., Wang, Y. B., Li, M., Yeung, D.-Y., 2022. Earthformer: Exploring space-time transformers for earth system forecasting. *Advances in Neural Information Processing Systems*, 35, 25390–25403.

Garioud, A., Gonthier, N., Landrieu, L., De Wit, A., Valette, M., Poupée, M., Giordano, S., Wattrelos, B., 2024. FLAIR: a Country-Scale Land Cover Semantic Segmentation Dataset From Multi-Source Optical Imagery.

Garioud, A., Gonthier, N., Landrieu, L., Wit, A. D., Valette, M., Poupée, M., Giordano, S., Wattrelos, B., 2023. Flair: a country-scale land cover semantic segmentation dataset from multi-source optical imagery.

Garioud, A., Peillet, S., Bookjans, E. M., Giordano, S., Wattrelos, B., 2022. FLAIR #1: semantic segmentation and domain adaptation dataset. *ArXiv*, abs/2211.12979. https://api.semanticscholar.org/CorpusID:253801541.

Garnot, V. S. F., Landrieu, L., 2020. Lightweight temporal self-attention for classifying satellite images time series. *Advanced Analytics and Learning on Temporal Data: 5th ECML PKDD Workshop, AALTD 2020, Ghent, Belgium, September 18, 2020, Revised Selected Papers 6*, Springer, 171–181.

Gehring, J., Auli, M., Grangier, D., Yarats, D., Dauphin, Y. N., 2017. Convolutional sequence to sequence learning. *International conference on machine learning*, PMLR, 1243–1252.

Giordano, S., Bailly, S., Landrieu, L., Chehata, N., 2018. Temporal Structured Classification of Sentinel 1 and 2 Time Series for Crop Type Mapping. https://hal.archives-ouvertes.fr/hal-01844619.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R., 2022. Masked autoencoders are scalable vision learners. *CVPR*, 16000–16009.

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

Ienco, D., Interdonato, R., Gaetano, R., Minh, D. H. T., 2019. Combining Sentinel-1 and Sentinel-2 Satellite Image Time Series for land cover mapping via a multi-source deep learning architecture. *ISPRS Journal of Photogrammetry and Remote Sensing*, 158, 11–22.

Kanyamahanga, H., Rottensteiner, F., 2024. Land cover classification based on multiscale time series of satellite and aerial images. *DGPF-Jahrestagung 2024 - Stadt, Land, Fluss - Daten vernetzen*, Geschäftsstelle der DGPF, 223–235.

Krizhevsky, A., Sutskever, I., Hinton, G. E., 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., Jackel, L. D., 1989. Backpropagation applied to handwritten ZIP code recognition. *Neural computation*, 1(4), 541–551.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021a. Swin transformer: Hierarchical vision transformer using shifted windows. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 9992–10002.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021b. Swin transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ma, Y., Li, Y., Feng, K., Xia, Y., Huang, Q., Zhang, H., Prieur, C., Licciardi, G., Malha, H., Chanussot, J. et al., 2021. The outcome of the 2021 IEEE GRSS data fusion contest-Track DSE: Detection of settlements without electricity. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14, 12375–12385.

Marmanis, D., Wegner, J. D., Galliani, S., Schindler, K., Datcu, M., Stilla, U., 2016. Semantic segmentation of aerial images with an ensemble of CNNs. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 3, 473–480.

Niu, R., Sun, X., Tian, Y., Diao, W., Chen, K., Fu, K., 2021. Hybrid multiple attention network for semantic segmentation in aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–18.

Ruder, S., 2016. An overview of gradient descent optimization algorithms. *CoRR*, abs/1609.04747. http://arxiv.org/abs/1609.04747.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. et al., 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115, 211–252.

Rußwurm, M., Körner, M., 2018. Convolutional LSTMs for cloud-robust segmentation of remote sensing imagery. *arXiv preprint arXiv:1811.02471*.

Tarasiou, M., Chavez, E., Zafeiriou, S., 2023. Vits for sits: Vision transformers for satellite image time series. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10418–10428.

Turkoglu, M. O., D’Aronco, S., Perich, G., Liebisch, F., Streit, C., Schindler, K., Wegner, J. D., 2021. Crop mapping from image time series: deep learning with multi-scale label hierarchies.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Voelsen, M., Lauble, S., Rottensteiner, F., Heipke, C., 2023. Transformer Models for Multi-Temporal Land Cover Classification Using Remote Sensing Images. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 10, 981–990.

Voelsen, M., Teimouri, M., Rottensteiner, F., Heipke, C., 2022. Investigating 2D and 3D convolutions for multitemporal land cover classification using remote sensing images. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 3, 271–279.

Wang, S., Gao, J., Li, Z., Zhang, X., Hu, W., 2023. A closer look at self-supervised lightweight vision transformers. *Proceedings of the 40th International Conference on Machine Learning*, Proceedings of Machine Learning Research, 202, PMLR, 35624–35641.

Yan, J., Liu, J., Liang, D., Wang, Y., Li, J., Wang, L., 2023. Semantic Segmentation of Land Cover in Urban Areas by Fusing Multisource Satellite Image Time Series. *IEEE Transactions on Geoscience and Remote Sensing*, 61, 1-15.

Yang, S., Chen, Q., Yuan, X., Liu, X., 2016. Adaptive coherency matrix estimation for polarimetric SAR imagery based on local heterogeneity coefficients. *IEEE Transactions on Geoscience and Remote Sensing*, 54(11), 6732–6745.

Yuan, Q., Shen, H., Li, T., Li, Z., Li, S., Jiang, Y., Xu, H., Tan, W., Yang, Q., Wang, J. et al., 2020. Deep learning in environmental remote sensing: Achievements and challenges. *Remote Sensing of Environment*, 241, 111716.

Yuan, X., Sarma, V., 2010. Automatic urban water-body detection and segmentation from sparse ALSM data via spatially constrained model-driven clustering. *IEEE Geoscience and Remote Sensing Letters*, 8(1), 73–77.