Classification of Satellite Image Time Series and Aerial Images Based on Multiscale Fusion and Multilevel Supervision

Hubert Kanyamahanga, Mareike Dorozynski, Franz Rottensteiner

Institute of Photogrammetry and GeoInformation, Leibniz Universität Hannover – Germany
(kanyamahanga, dorozynski, rottensteiner)@ipi.uni-hannover.de

Keywords: Land Cover Classification, Transformers, FCN, Multi-stage Supervision, Multiscale Data Fusion

Abstract

A large variety of sensors can be used for monitoring processes on the Earth's surface. Different sensors can capture complementary information of the same observed region. For instance, aerial images offer a high spatial resolution but at a low temporal resolution, whereas satellite image time series (SITS) capture temporal variations with a high repetition rate, e.g. seasonal changes, but with limited spatial resolution. This paper presents a method to jointly exploit the strengths of SITS and aerial images for land cover classification. In this context, it is a challenge to train a classifier given the large difference in resolutions. We utilise convolutions to extract spatial information and consider self-attention in the temporal dimension for SITS. Additionally, a multi-resolution supervision strategy is proposed, applying auxiliary losses at different stages of the SITS decoder to enhance feature learning. Features extracted from SITS data are fused via a cross attention module with features determined from aerial images at the same spatial resolution by a SegFormer network before predicting land cover at the geometrical resolution of the aerial image. We perform comparative experiments on an existing benchmark dataset, showing that the convolution- and attention-based fusion of a SITS from Sentinel-2 with aerial image improves the classification results by +1.9% in the *mean IoU* and +2% in the *OA* compared to a method based on aerial images only.

1. INTRODUCTION

Recent developments in remote sensing (RS) have significantly encouraged the use of multi-sensor data for applications such as land cover classification, i.e. the task of assigning a class label representing the physical material of the Earth's surface to each pixel of an image. Multiple sensors can be used to acquire data with complementary information about the same observed region. For example, aerial imagery delivers textural information at very high geometrical resolution, but usually with high revisit times, so that there is no information about the changing appearance of objects during the vegetation cycle. On the other hand, satellite systems, such as Sentinel-2, have short revisit times, so that the resultant images can capture temporal changes, but usually at a coarser spatial resolution, e.g. with a ground sampling distance (GSD) of 10 m or more. This causes problems in detecting smaller objects. Thus, it is interesting to develop methods for combining such data for improved land cover classification.

For several years, deep learning methods have been exploited to process RS data. For aerial and satellite images of a single epoch, Fully Convolutional Networks (FCNs) with encoder-decoder architectures such as U-Net (Ronneberger et al., 2015) are frequently used for land cover classification. For SITS data, methods such as 3D-Convolutional Neural Networks (CNNs) (Li et al., 2022) or Recurrent Neural Networks (RNNs) (Sharma et al., 2018) have been used to extract spatial and temporal information. Methods based on 3D-CNNs consider time as an additional dimension of the input data and learn 3D filter kernels for a convolution in the spatial and temporal dimensions (Ji et al., 2018; Li et al., 2022). RNNs are designed for sequential data and capture temporal dependencies of the time series while processing one image at a time, maintaining a memory of previous inputs and generating the output based on the memory and the current input (Rußwurm and Körner, 2017). Recently, they have been challenged by vision transformers (ViTs) (Dosovitskiy et al., 2021), which have been adapted to capture both spatial and temporal long-range dependencies in SITS data, yielding promising results in pixel-wise classification (Tarasiou et al., 2023; MacDonald et al., 2024; Voelsen et al., 2024). However, the methods cited so far use a single input modality only. They also convert the input images into sequences of patches, resulting in a critical loss of spatial details for SITS.

The fusion of multi-source data has emerged as an approach to improve the classification accuracy beyond what can be achieved using single data modalities. Fusion methods can be classified into early, mid, late and decision fusion (Garnot et al., 2022), depending on the stage at which the fusion of multimodal input is performed. In early fusion, the raw features from various sources are concatenated before being processed by the network. In mid-level fusion (Garioud et al., 2024), features are extracted from each modality before being combined. This enables interaction between modalities, creating richer, joint representations in subsequent network layers. In late fusion, the integration is performed at high-level stages, after independent feature extraction from each modality. This is different from decision level fusion, in which each modality is processed independently and the final decision is determined based on the most confident predictions. In all of these approaches, the fusion itself involves either element-wise addition or channel-wise concatenation before classification. Recently, attention-based methods have been adopted, allowing one modality to weight the features of another. Ren et al. (2024) use a cross-attention approach to fuse optical and synthetic aperture radar (SAR) images, which improved the classification accuracy compared to element-wise addition fusion. However, to the best of our knowledge, attention-based fusion approaches have not been fully exploited to fuse aerial and Sentinel-2 images.

A common challenge in training multi-source classifiers arises

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-G-2025-477-2025

when most discriminative information is consolidated in one modality, leading to less relevant features and weaker predictions from the other one. Consequently, multi-modal fusion tends to prioritize the stronger modality, leading to a reduced supervision signal for the weaker one. This can limit the network from fully exploiting inter-modal relationships, potentially limiting the performance of the multi-modal predictions. To address this issue, a separate loss can be used to monitor each modality. Existing approaches (Ienco et al., 2019; Benedetti et al., 2018) add auxiliary losses at the last layer of the network to the main loss function to support the supervision of each modality independently. Although this has been shown to improve the overall network performance, the supervision at various feature resolutions has not been investigated yet. We hypothesize that multiscale auxiliary supervision improves the semantical meaningfulness of features at earlier stages in the network and thus, improves land cover classification.

This paper presents a hybrid convolution- and attention-based method with the goal to learn a joint representation of aerial and co-registered SITS data to obtain a land cover map at the GSD of the aerial image. We hypothesize that, for low-resolution SITS, convolutions may learn spatial information more effectively, and attentions can be more useful for extracting temporal information. We present an end-to-end learning procedure that in principle can deal with all types of multiscale RS data. This is achieved by designing a two-branch architecture which learns sensor specific properties independently. The combination of the features from the aerial and SITS branches is based on a fusion module which enables mutual information exchange between the two modalities. Additionally, we introduce auxiliary loss functions at multiple resolutions of the SITS branch decoder to allow the SITS branch network to capture semantically meaningful features at multiple scales. We train and evaluate our method on an existing benchmark dataset (Garioud et al., 2024), showing the benefits of introducing multi-temporal information from Sentinel-2 compared to uni-modal aerial image classification. The scientific contributions of this paper can be formulated as follows:

- We propose a new method for the fusion of aerial images and SITS of varying length for land cover classification based on convolutions for spatial feature extraction and using self-attention in the temporal dimension.
- We introduce intermediate auxiliary losses at various levels of the SITS decoder branch to help the network to learn to capture features at different resolutions.
- We introduce a fusion module based on cross-attention to facilitate the exchange of complementary information between aerial and SITS data.
- We present the results of extensive experiments to evaluate our method, assess the impact of various components on the quality of the results, and compare our method to baselines from the literature.

2. RELATED WORK

We start this review by discussing related work that uses CNNs to integrate and fuse multiscale data for pixel-wise classification. Next, we review transformer models for semantic segmentation of RS data, particularly SITS, and we discuss existing multiscale fusion approaches. Finally, we review approaches for multiscale supervision.

CNN-based models for multiscale data: The integration of multiscale RS data has been investigated in several works. Benedetti et al. (2018) use a Gated Recurrent Unit network to process Sentinel-2 time series (10 m GSD) and a 2D-CNN branch to extract features from mono-temporal SPOT-6 images (2 m GSD). The resultant features are concatenated and provided to a decoder to predict land cover at the GSD of SPOT-6. Gbodjo et al. (2021) combine Sentinel-2 and SPOT-6 data with SITS from Sentinel-1. The Sentinel-1 and SPOT data are processed by 2D-CNN encoders, while the Sentinel-2 data are analyzed in an encoder applying a convolution only in the temporal dimension. The resultant features are also concatenated and used to predict land cover at the GSD of SPOT-6. Both approaches are limited to a fixed number of timesteps. Also, the difference in the GSDs is relatively small (10 m vs. 2 m).

There are not many CNN-based approaches that combine SITS with aerial data. Bergamasco et al. (2023) use a 3D-CNN to extract spatial and temporal features from Sentinel-2 SITS, concatenating them with features extracted from aerial images (0.2 m GSD) using a residual network. After reducing the number of features by a 1×1 convolution, they are fed to a decoder to classify different types of pasture at the GSD of the aerial images. The combined method is shown to achieve better classification results compared to an uni-modal approach. However, the method has limitations in differentiating classes that are semantically similar. This can be due to a small temporal receptive field in 3D-CNN, which is determined by the depth of the network and the size of the convolutional filters. Though increasing network depth expands the temporal receptive field, 3D-CNNs may still struggle with long-range dependencies, as each layer captures only local information.

Attention-based models for multiscale data: Transformer models (Vaswani et al., 2017) are based on self-attention modules to model long-range dependencies in input sequences. They have been adapted to various applications in computer vision (Dosovitskiy et al., 2021; Liu et al., 2021; Strudel et al., 2021; Wang et al., 2021; Xie et al., 2021). Compared to 3D-CNNs, typically requiring fixed input dimensions, they can cope with varying sequence lengths. Several works have adapted attention-based models to the extraction of spatial and temporal information from SITS (Garnot et al., 2020; Tarasiou et al., 2023; MacDonald et al., 2024; Voelsen et al., 2024). Voelsen et al. (2024) extended the Swin transformer (Liu et al., 2021) for processing SITS. For each image of the SITS, Swin transformer blocks are executed in parallel to extract spatial features, and the outputs are processed jointly by a temporal attention block. The modified Swin transformer outperforms purely CNN-based models for the task of generating multi-temporal land cover maps from Sentinel-2 time series. Tarasiou et al. (2023) adapted the ViT (Dosovitskiy et al., 2021) for crop classification based on SITS data. They first compute attentions between all timesteps of corresponding patches at the same spatial location. After that, the outputs are reshaped and the attentions are computed between all patches of the same timestep. Though this model was shown to achieve better results with fewer parameters compared to hybrid convolution-attention approaches (Garnot and Landrieu, 2021), it has a quadratic complexity w.r.t. to the input size, leading to higher hardware demand when working with larger inputs. Neither Voelsen et al. (2024) nor Tarasiou et al. (2023) combine multiple modalities at different GSDs.

Very few approaches have used attention-based models to combine SITS data with aerial images. Garioud et al. (2024) pro-

posed a two-branch U-Net-based architecture to fuse Sentinel-2 SITS with aerial images. They adopt the U-TAE model of Garnot and Landrieu (2021), a modified U-Net with a Temporal self-Attention Encoder (TAE), to extract temporal information from a SITS. The aerial images are processed by a U-Net to produce pixel-wise class predictions; in order to fuse the two modalities, the SITS features are added element-wise to the encoder features of all levels in the skip connections. Temporal attention is only considered at the lowest resolution and the resultant features are upsampled to higher resolutions. This might not fully capture temporal variations at different spatial resolution levels. Moreover, the element-wise addition used for fusion in (Garioud et al., 2024) might not be optimal in situations in which certain modalities are more relevant for specific land cover types. Kanyamahanga and Rottensteiner (2024) extend (Garioud et al., 2024) by introducing transformers from (Tarasiou et al., 2023; Voelsen et al., 2024) into the SITS branch. Similarly, Heidarianbaei et al. (2024) introduced a variant of ViT model to simultaneously extract spatial and temporal information from SITS data. They show that using transformer models to learn spatial and temporal information barely improves the classification. Their approaches also rely on patchification of the SITS, which could be problematic given the low resolution of these data.

Deep multiscale network supervision: A typical way to train a deep neural network involves minimizing a loss function, measuring the discrepancy between the output of the last layer of the network and reference labels. For multiscale fusion methods, it is preferred to have a separate loss function for each modality, training the network by a joint loss that is a linear combination of the individual terms (Garioud et al., 2024). Some works add auxiliary loss terms to each network branch on the top of the main objective loss to boost the prediction performance of each branch prior to final classification (Benedetti et al., 2018; Ienco et al., 2019). Again, the supervision is applied at the last layer of the network; this does not allow the intermediate layers to learn how to refine the features. Nevertheless, this is assumed to be useful for complex land cover types where different features may occur at different spatial resolutions.

The approach presented in this work aims to address the aforementioned limitations of existing approaches for the integration of SITS and aerial image for pixel-wise classification (Bergamasco et al., 2023; Garioud et al., 2024; Kanyamahanga and Rottensteiner, 2024; Heidarianbaei et al., 2024). Similar to (Voelsen et al., 2024), we use both convolutional and attention-based models to extract spatial and temporal information from the SITS data, but we combine this information with high-resolution features extracted from an aerial image (Xie et al., 2021). We introduce auxiliary losses at different stages to be able to exploit the multi-temporal information contained in the SITS data across multiple resolutions in a better way. The multiscale classification method proposed in this work is designed to solve complex land cover classification tasks, especially when the difference in the GSDs of the given imagery is large (e.g., 10 m vs. 0.2 m).

3. METHODOLOGY

The goal of our method is to exploit the complementary strengths of SITS and aerial data acquired over the same area to predict the land cover of the depicted scene at a pixel-level at the GSD of the aerial image. For that purpose, a network architecture consisting of two branches is proposed: The *SITS branch* is designed for extracting features from the given SITS, whereas the *aerial branch* extracts features from the aerial image. Inspired by (Voelsen et al., 2024), the SITS branch relies on convolutions for spatial feature extraction and considers self-attention in the temporal dimension. However, unlike (Voelsen et al., 2024), our method computes temporal attentions at a pixel-level, mitigating the problem of the patchification process of transformer models (Dosovitskiy et al., 2021) which can lead to a loss of spatial detail. In addition, we introduce a new multi-resolution supervision approach in the SITS branch, adding auxiliary loss terms at different stages of that branch. For the aerial branch, we use the SegFormer network (Xie et al., 2021), which has shown to be efficient in learning multiscale features with a small number of parameters. We introduce a cross-attention approach for fusing features extracted from the SITS branch with features determined in the aerial branch, allowing the network to learn on which features to focus in the classification. Details about the network architecture and the training procedure are given in the subsequent sections.

3.1 Network Architecture

An overview of our proposed network architecture is presented in Figure 1. The input consists of a georeferenced SITS X^S with T timesteps, each image having C^S spectral bands and covering an area of $H^S \times W^S$ pixels at the GSD of the SITS, along with an aerial image X^A with C^A spectral bands and covering an area of $H^A \times W^A$ pixels at a higher spatial resolution. The area covered by the aerial image corresponds to a subset of the area covered by the SITS. The output is a land cover map of dimension $H^A \times W^A$ at the GSD of the aerial image.

3.1.1 SITS Branch: The SITS branch uses an encoder-decoder architecture (cf. Figure 2). The encoder is based on a modified version of (Voelsen et al., 2024), which uses the Swin Transformer (Liu et al., 2021) and processes patches of 4×4 pixels. To avoid this patchification, which reduces the spatial resolution of the SITS, we remove the patch partitioning step and use convolutions for spatial feature extraction while considering self-attention in the temporal dimension at pixel level. We refer to this encoder as T-ConvFormer. Its output is processed by a decoder based on UPerNet (Xiao et al., 2018).

The input to the T-ConvFormer encoder is a SITS organized into a four-dimensional tensor of shape $T \times C^S \times H^S \times W^S$. The encoder extracts spatial and temporal information from the given SITS at every position (h^S, w^S). It consists of four processing stages, each generating a feature map at a different resolution (feature maps F_i with dimensions $C_i^S \times H^S/2^{i-1} \times W^S/2^{i-1}$, i representing the stage index). One stage consists of several spatial temporal blocks with convolution and attention (STB-CA) as introduced in (Voelsen et al., 2024). Stage 1 consists of two such blocks (cf. Figure 2); the output F_1 is passed to the decoder before being downsampled to serve as an input for encoder in stage 2. In the subsequent stages, multiple STB-CA blocks are applied (two in stages 2 and 4, six in stage 3) to extract multiscale features from the SITS.

The components of the STB-CA blocks are shown in Figure 3. First, convolutions are executed in parallel for each time step to extract spatial features (violet boxes in Figure 3). These feature maps are stacked along the temporal dimension before being processed by a temporal attention block only considering attention in the temporal domain (red box in Figure 3). As no patchification is performed, at stage 1 the temporal attention is

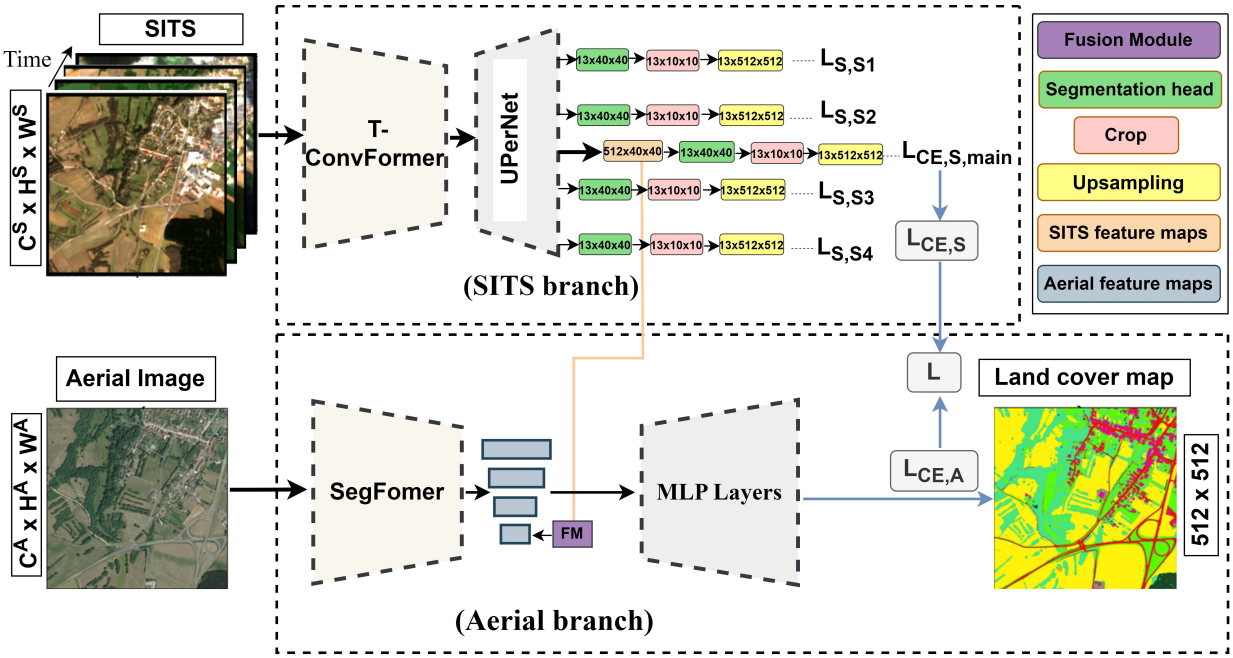

Figure 1. Network architecture for the joint classification of SITS and an aerial image. The SITS branch encoder uses convolutions in the spatial dimension and self-attention in the temporal dimension. The aerial branch processes the aerial image using a SegFormer encoder. The SITS feature map is cropped and upsampled to the lowest resolution feature map of the aerial branch and the SITS and aerial features are combined in the fusion module (FM). A decoder, consisting of a stack of MLP layers, uses the combined features to predict the land cover map. In training, a classification loss is minimized for both the SITS and the aerial branches, with auxiliary losses added at various stages of the SITS branch. At test time, no labels are predicted for the SITS.

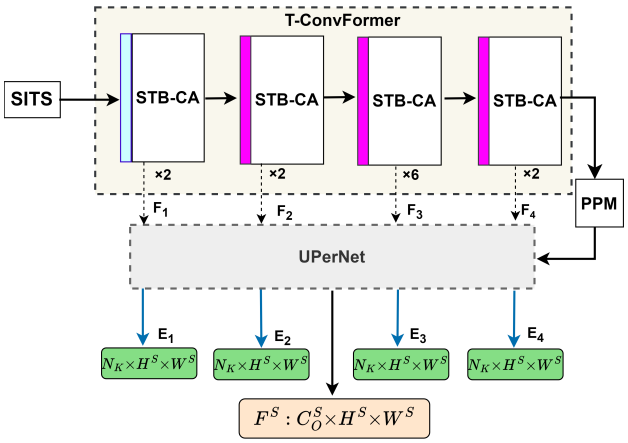

Figure 2. The SITS branch. The feature maps F_1, F_2, F_3, F_4 produced by T-ConvFormer are used by UPerNet to generate a feature map F^S of size $C_O^S \times H^S \times W^S$ that is integrated into the aerial branch. PPM: Pyramid Pooling Module, light blue rectangle: linear embedding, blue arrows: bilinear upsampling by a factor of 2, magenta rectangles: down-sampling, green boxes: softmax outputs. STB-CA as in Figure 3.

applied at pixel level at the original resolution of the SITS. The temporal dimension is considered in all stages. After stage 4, the temporal dimension is discarded by computing the average of the features of each time step for each pixel to achieve the same extent in temporal dimension like the uni-temporal aerial features for the fusion module (details in Section 3.1.3).

The feature maps F_i produced by the four stages of the encoder are mapped to a feature map F^S of dimensions $C_O^S \times H^S \times W^S$ using the UPerNet (Xiao et al., 2018) decoder (yellow-orange box in Figure 2). F^S represents the multi-temporal information and has the same spatial resolution as the input X^S. In training, the lower-resolution features from UPerNet are upsampled to the spatial resolution of the input X^S, and a softmax layer is applied to the upsampled features to generate pixel-wise class scores $E_i : N_K \times H^S \times W^S$ (green boxes in Figure 2), used to compute auxiliary losses (cf. Section 3.2).

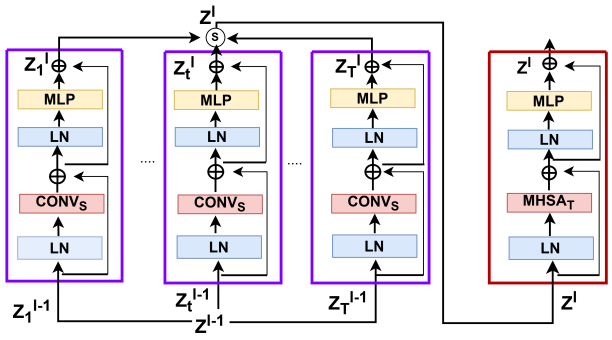

Figure 3. Spatio-temporal block $STB-CA$ adapted from (Voelsen et al., 2024): parallel blocks (violet boxes) extract spatial features for all timesteps based on convolutions; this is followed by a temporal attention block (red box). All of these sub-blocks form one $STB-CA$ block l. Z^{l-1} and Z^l: input and output to the $STB-CA$ module l, respectively. *LN*: layer normalization, CONV_S: depthwise separable convolutions (3 x 3 depthwise conv., 1 x 1 pointwise conv.). MHSA_T: multi-head self attention in the temporal domain, *MLP*: Multilayer Perceptron, +: element-wise addition, S: stacking of outputs.

3.1.2 Aerial Branch: This branch combines a SegFormer encoder and a MLP-based decoder, both from (Xie et al., 2021). Before the SegFormer output is processed by the decoder, our new fusion module (cf. Section 3.1.3; FM in Figure 1) integrates the output of the SITS branch with the feature map generated by SegFormer having the lowest spatial resolution. The input of the aerial branch is an image organized into a tensor of shape $C^A \times H^A \times W^A$, and the output is a pixel-wise land cover map $N_K \times H^A \times W^A$.

The SegFormer encoder is presented in Figure 4. It consists of four processing stages. In stage 1, the aerial image is divided into overlapping patches of size 7 × 7, where each patch overlaps with its adjacent patches by 3 pixels. These patches are flattened into 1D vectors and are linearly projected to high-dimensional vector embeddings. Two transformer blocks are applied to compute spatial attentions between all embeddings, resulting in a feature map G_1 which is downsampled to form the input of stage 2. In the subsequent stages, two transformer blocks are applied per stage. Each of them includes*efficient self-attention* layers (Efficient-Self-Att in Figure 4), reducing the computational complexity compared to standard self-attention (Xie et al., 2021). Instead of positional encodings, SegFormer combines 3 × 3 convolutions with MLPs (Mix-FFN in Figure 4) to encode the positional information of each patch. We direct the reader's attention to (Xie et al., 2021) for more details about SegFormer. The output of each stage is a three-dimensional tensor G_j: $C_j^A \times H^A/2^{j+1} \times W^S/2^{j+1}$), where $j \in \{1, 2, 3, 4\}$ denotes the stages of the SegFormer.

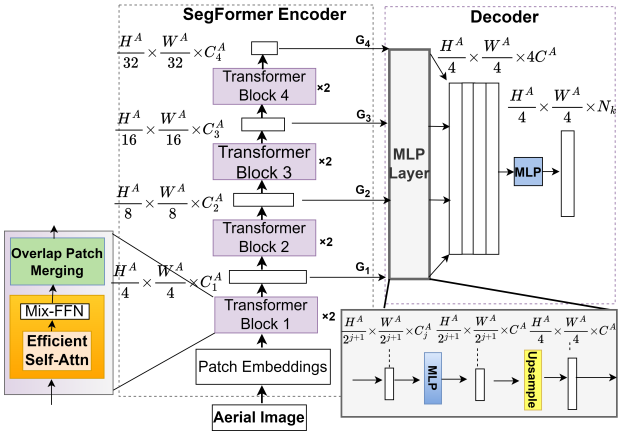

Figure 4. Architecture of the aerial branch. It produces four feature maps G_1 - G_4; the fusion of G_4 with the SITS features in the fusion module is not shown for simplicity. The resultant features are provided to a decoder to produce a land cover map. Efficient Self-Attn combines standard self-attention with a sequence reduction process (Wang et al., 2021). Mix-FFN: Mix-feed forward network combining 3 × 3 convolutions with a multilayer perceptron (MLP) to encode the spatial position of each patch. N_K: number of classes.

The outputs of the SegFormer encoder are fused with features from the SITS branch via a cross-attention layer, as described in Section 3.1.3. The fused features serve as inputs to a decoder, consisting of a stack of MLP layers. First, a MLP is used to reduce the channel dimension of the four feature maps. These feature maps are upsampled to the same spatial dimension ($C^A \times H^A/4 \times W^A/4$, i.e. 4 times the GSD of the aerial image) by bilinear interpolation before being concatenated. Next, a MLP layer is used to fuse the concatenated feature maps. Finally, another MLP takes the fused feature maps to predict the class scores, on the basis of which the land use map is generated.

3.1.3 Fusion with cross-attention: The features F^S extracted from the SITS are fused with the feature map G_4, i.e. the one with the lowest resolution determined in the aerial branch. Fusion is based on the standard multi-head cross-attention (Vaswani et al., 2017). It is applied two times to generate new SITS ($\mathbf{WF}_S$) and aerial ($\mathbf{WG}_A$) features:

$$\begin{aligned} \mathbf{WF}_S &= MHA\left(\mathbf{Q}_A, \mathbf{K}_S, \mathbf{V}_S\right) \\ \mathbf{WG}_A &= MHA\left(\mathbf{Q}_S, \mathbf{K}_A, \mathbf{V}_A\right), \end{aligned} \tag{1}$$

where MHA denotes multi-head attention according to (Vaswani et al., 2017) and $\mathbf{Q}_S$, $\mathbf{K}_S$, $\mathbf{V}_S$ and $\mathbf{Q}_A$, $\mathbf{K}_A$, $\mathbf{V}_A$ are the query, key and value matrices generated by a linear projection from the SITS and aerial features, respectively. Note that in the cross-attention for computing new features for one of the modalities, this modality is used to generate the key and value matrices, whereas the query matrix is generated from the other modality. The new features $\mathbf{WG}_A$ and $\mathbf{WF}_S$ are concatenated and a 1 × 1 convolution $CONV_{1\times1}$ is applied to reduce the channel dimension:

$$\mathbf{M}_4 = CONV_{1\times1}(Concat(\mathbf{WG}_A, \mathbf{WF}_S)). \tag{2}$$

The combined features ($\mathbf{M}_4$), which now contain both spatial and multi-temporal information from the aerial image and the SITS; they are substituted for the lowest-resolution feature map G_4 of the aerial branch before being presented to the decoder.

3.2 Training with main and auxiliary losses

Training is based on minimizing a loss function consisting of one term per branch. We additionally introduce intermediate auxiliary losses ($L_{S,Si}$, one term for each decoder level of the SITS branch) to support the generation of meaningful features during network training. The weighted sum of those auxiliary losses is added to the main loss ($L_{S,main}$) monitoring the SITS branch. By applying supervision at multiple levels, we try to support the SITS network to learn more discriminative features and, thus, ultimately to provide more useful information for the classification task. All losses are based on the categorical Cross Entropy (CE) loss:

$$L_{CE} = -\sum_{v=1}^{N_P}\sum_{u=1}^{N_K} t_{uv}\log(p_{uv}), \tag{3}$$

where N_P is the number of pixels of the output label image, v is the index of a pixel, N_K is the number of classes, and u is the index of a specific class. The indicator variable t_{uv} indicates whether the reference class label of pixel v is u ($t_{uv} = 1$) or not ($t_{uv} = 0$), and p_{uv} is the softmax output for pixel v to correspond to class u. The loss for the SITS branch becomes

$$L_{CE,S} = \alpha_0 \cdot L_{CE,S,main} + \sum_{i=1}^{N_S} \alpha_i \cdot L_{CE,S,Si}, \tag{4}$$

where $L_{CE,S,main}$ is the cross-entropy loss defined in equation 3 applied to the output of a softmax layer used to predict class scores from the feature map F^S and $L_{CE,S,Si}$ are auxiliary CE loss terms applied to different auxiliary softmax outputs E_i of the SITS branch (cf. Section 3.1.1). N_S denotes the number of stages of the SITS branch. The weights α_0 and α_i

modulate the influence of the respective losses on the training procedure. Denoting the CE loss computed on the basis of the output of the aerial branch by $L_{CE,A}$, the total loss is

$$L = \lambda_S \cdot L_{CE,S} + \lambda_A \cdot L_{CE,A}, \quad (5)$$

where $L_{CE,S}$ is defined according to equation 4 and λ_S and λ_A are weights of the two loss terms. To compute the auxiliary losses ($L_{CE,S,Si}$) as well as the main loss ($L_{CE,S,main}$), the SITS branch has to predict class probabilities. Thus, a softmax layer is applied to each feature map generated by that branch when the network is trained. As a reference is only provided for the aerial image, which covers a smaller area than the SITS, the class scores are cropped to the area of overlap. In order to calculate the $L_{CE,S,main}$, the cropped class scores are upsampled to the spatial resolution of the aerial image. To compute the auxiliary losses ($L_{CE,S,Si}$), the reference labels are downsampled to the GSD of SITS using majority vote. The loss in equation 5 is minimized using the Adam optimizer (Kingma and Ba, 2015). For the SITS branch, the network weights are randomly initialized based on (He et al., 2015), while the parameters of the aerial branch are initialized by weights pre-trained on ImageNet (Xie et al., 2021).

4. EXPERIMENTS

4.1 Test Dataset

In our experiments, we use the French Land cover from Aerospace ImageRy (FLAIR) #2 Challenge dataset (Garioud et al., 2024), consisting of mono-temporal multispectral aerial image and height data acquired between 04/2018 and 11/2021 and SITS acquired by Sentinel-2 over a period of one year in France. The dataset contains imagery and reference data from 916 areas in France, with a total area of about 817 km^2. All images and label maps are georeferenced in the same coordinate system. The aerial images have 4 channels (RGB, near infrared) at a GSD of 20 cm. A normalized digital surface model is available as an additional input band, thus C^A = 5. The SITS data consist of Sentinel-2 L2A images containing bottom-of-atmosphere reflectance values, and cloud and snow masks (Drusch et al., 2012). We use C^S = 10 channels at a GSD of 10 m, upsampling the six bands with a GSD of 20 m by nearest neighbour resampling. Images having more than 5% of cloud cover according to the cloud masks are eliminated, so that the number of satellite images per test varies between 20 and 110. We follow the procedure used in (Garioud et al., 2024), which requires a fixed-length input, and preprocess the SITS by computing monthly average reflections considering cloud-free pixels in the satellite images, so that the maximum number of timesteps available for an area is 12. However, the number of timesteps might vary because there are months for which there is not a single cloud-free image of a test area. There is a pixel-wise reference at the GSD of aerial images which differentiates 13 land cover classes: *building (bld.)*, *pervious surface (pvs.)*, *impervious surface (ips.)*, *bare soil (bs.)*, *water (wt.)*, *coniferous (cfs.)*, *decidous (dcs.)*, *brushwood (bsd.)*, *vineyard (vyd.)*, *herbaceous vegetation (hvg.)*, *agricultural land (agr.)*, *plowed land (pld.)* and *other*. The class *other* corresponds to unknown land cover. The class distribution is very imbalanced, with class frequencies varying between 1.1% (*other*) and 19.8% (*hvg*).

Each area is split into subsets (referred to as tiles) covering 512 x 512 pixels at the GSD of the aerial image. The SITS of each tile is sampled so that it covers a larger area than aerial with aerial patch in the center, resulting in a size of 40×40 pixels at the GSD of 10 m. Altogether there are 77,762 tiles, each with an aerial image, a SITS (with the number T of timesteps varying between 1 and 12) and a reference label map. Garioud et al. (2024) defined a training set consisting of 61,712 tiles and a test set consisting of the remaining 16,050 tiles. More details can be found in (Garioud et al., 2024). We use the same definition, further splitting the training set into a set of 48,812 tiles to be used for updating the parameters (we will call this set training set in the rest of the paper) and a validation set consisting of 12,900 tiles.

4.2 Experimental Protocol

We apply the methods described in Section 3 to the data described in 4.1, where the input dimensions for the aerial and the SITS data are $H^A \times W^A = 512$ and $H^S \times W^S = 40$, respectively. The patch size for the tokens in the SegFormer model for aerial branch of the network is set to [7, 4, 4, 4] for the four stages (Xie et al., 2021). The training procedure is described in Section 3.2. In training, we also applied data augmentation, using random rotations by 90°, 180°, 270°, horizontal and vertical flipping. Training is carried out for a maximum of 100 epochs, but training is stopped if the validation accuracy does not increase for 15 epochs (early stopping), which is the case before the maximum of 100 epochs is reached. We used the Adam optimizer (Kingma and Ba, 2015) with the parameters β_1 = 0.9 and β_2 = 0.999. The batch size is set to 4, the learning rate is set to $6e^{-5}$ and is decreased by a factor of 0.7 every 10 epochs. The weights for the loss terms in equations 4 and 5 are set to α_0 = 0.6, α_i = 0.4, λ_A = 0.7 and λ_S = 0.3. We selected these values based on their best performance on the validation dataset. Training is carried out on a cluster with two NVIDIA A100 80GB GPUs. All the models are implemented using the PyTorch Lightning Framework.

We conducted several sets of experiments, comparing our method to five other methods and performing an ablation study with respect to the components of our method (cf. Table 1). The first baseline method (referred to as U-Net) uses the U-Net of (Garioud et al., 2022) to predict land cover only based on the aerial image. This is also true for the second baseline, SegFormer (Strudel et al., 2021); the results from these two methods are compared with those of methods also using SITS to assess the contribution of the SITS to the classification accuracy. The third method (U-T&T) integrates the areal image with a SITS based on (Garioud et al., 2024). The fourth and fifth method (Swin and TSViT, respectively) are described in (Kanyamahanga and Rottensteiner, 2024). They both integrate SITS, using U-Net (Garioud et al., 2022) for the aerial branch. Our own method, described in Section 3, is referred to as T-ConvFormer. We also conduct an ablation study to assess the impact of different components on the design choices of our approach. We compare the results of T-ConvFormer with two of its variants. The first one, referred to as T-CF-0, uses element-wise addition of features instead of the cross-attention fusion module, and it does not consider the auxiliary losses in training. The second one (T-CF-1) is based on cross-attention fusion, but still does not consider the auxiliary losses.

Each experiment was repeated three times, each time starting from a different random initialization of the weights and using random shuffling for batches. The classification results on the test images are compared to the reference. We report the intersection over union (IoU) for each class k:

Name	Aerial	SITS	CAF	AUX
U-Net	U-Net	✘	✘	✘
SegFormer	SegFormer	✘	✘	✘
U-T&T	U-Net	U-TAE	✘	✘
TSViT	U-Net	TSViT	✘	✘
Swin	U-Net	Swin	✘	✘
T-CF-0	SegFormer	T-FC	✘	✘
T-CF-1	SegFormer	T-FC	✔	✘
T-ConvFormer	SegFormer	T-FC	✔	✔

Table 1. Overview of the experiments conducted to evaluate our method and compare it to the state of the art. Name: name by which an experiment is referred to. Aerial / SITS: architectures used for processing the aerial image and the SITS; please refer to the main text for a description of methods and acronyms. Our method based on the T-ConvFormer encoder is denoted by T-CF. Columns CAF and AUX indicate whether the cross-attention fusion module and auxiliary supervision were applied.

$$IoU_k = \frac{TP_K}{TP_k + FP_k + FN_k}, \tag{6}$$

where TP_k, FP_k, and FN_k, denote the number of pixels that are true positives, false positives and false negatives, respectively, for a class *k*. The mean intersection over union (*mIoU*) is also computed by taking the average of the IoU_k values for all classes except *other*, following the protocol in (Garioud et al., 2024). We also compute the overall accuracy (*OA*), i.e. the proportion of correctly classified pixels. For these two compound metrics, we report the average and the standard deviations over the three test runs. We also report the inference time (IT) per aerial image tile in milliseconds. The FLAIR #2 challenge required this inference time of any approach to remain within 2.5 times the one of the U-T&T, claiming that this was a prerequisite for practical relevance (Garioud et al., 2024).

5. RESULTS

A summary of the average quality metrics achieved by all trained methods is presented in Table 2. As expected, the use of SITS data as an additional source of information leads to an increase in the overall performance w.r.t to the *mIoU* and *OA*. The U-Net, which uses only aerial images, achieved the lowest *mIoU* of 55.2%. Introducing SegFormer, another model which also uses aerial images, leads to a significant increase of 3% in the *mIoU* (55.2% vs 58.2%). By considering SITS, the U-T&T (Garioud et al., 2024), slightly improves the *mIoU* (56.8% vs 55.2%). The method based on Swin transformer model (Swin) achieved similar performance. The TSViT model performs slightly better, with a 0.2% improvement over the U-T&T model. Our method T-ConvFormer achieved the best *mIoU* score of 60.1% and complies with the computational cost constraint of FLAIR #2 competition, with an inference time being about 40% larger than the one of U-T&T. It can be seen that using SITS and combining convolutions and attention-based methods in the corresponding network branch leads to a significant improvement in the *mIoU* (+4.9%) compared to a U-Net model using a single-date aerial data, but also of +1.9% compared to SegFormer, which otherwise outperforms the compared methods based on SITS. Our method also shows an improvement of about +3% in *mIoU* over transformer-based approaches for processing SITS. T-ConvFormer also outperforms all other methods by a margin of about 2-3% in OA. This achievement can be attributed to various components introduced by our method which includes the attention based fusion and the multiscale supervision in leveraging the complementary strengths of multi-temporal information from SITS and spatial details from aerial images, as will be analysed below.

Method	**mIoU [%]**	**OA [%]**	**IT [ms]**
U-Net	55.2 ± 0.0	71.3 ± 0.0	22.6
SegFormer	58.2 ± 0.0	72.3 ± 0.0	42.3
U-T&T	56.8 ± 0.7	71.7 ± 0.0	64.7
Swin	56.9 ± 1.1	72.1 ± 0.1	45.6
TSViT	57.0 ± 0.0	71.9 ± 0.0	80.0
T-ConvFormer	**60.1** ± 0.0	**74.3** ± 0.0	90.0

Table 2. Mean IoU (mIoU) and Overall Accuracy (OA) for land cover classification [%] on the test set of the FLAIR #2 dataset achieved by different approaches. The numbers are the averages and standard deviations achieved in three independent test runs. Column Method gives the name of a method as defined in Table 1. IT: inference time in milliseconds [ms] per aerial image tile. Best results are indicated in **bold**.

Table 3 presents the class-specific *IoU* scores achieved on the test set of the FLAIR #2 dataset by all compared method. The numbers presented in the table do not show a very clear trend, but they do show that our method (T-ConvFormer) performs best on nine out of twelve classes, in case of *bs* by a large margin (+5.6% compared to the second best method). For the other eight classes, the margin by which our method outperforms the others is in the order of 1%-4% across classes. The only class for which our method is outperformed by a relatively large margin is *hvg*, and even more astonishingly, that class is differentiated best by U-Net, which achieved a relatively weak performance for most other classes. This might be the case because, despite being a vegetation class, *hvg* (e.g., parks or sport fields) does not show strong temporal variations compared to other vegetation classes, making SITS less beneficial. These results indicate that the impact of using SITS cannot be described simply as leading to an improvement for the vegetation classes.

Table 3 also indicates that some classes are easier to differentiate than others. The class frequency in the dataset appears to play a significant role, as some of the highest *IoU* scores (57%-88%) are obtained for classes that occur frequently (e.g., *ips*, *dcs*, *bld*, and *agr*), whereas some of the lowest scores are associated classes occurring less frequently (e.g., *bsd*, *pld*). In contrast, *hvg*, a class with a high frequency of occurrence, achieves a very low *IoU* score, while *vyd*, an underrepresented class, achieves a relatively high score. The poor performance for *hvg* may be attributed the high similarity between the appearance of some classes. For instance, distinguishing herbaceous vegetation areas (e.g. gardens, public parks, sport fields) from agricultural land, which partly corresponds to pastures, may be challenging.

Figures 5 and 6 show some qualitative results of the four approaches considering SITS for areas of different characteristics (urban and rural, respectively). The figures indicate that in these areas *agr*, *vyd*, and *cfs* are classified better by our approach (T-ConvFormer) (cf. the regions in the black circles). Overall, all compared models show similar performance on classes such as *bld*, *ips*, and *pvs*, which are not affected by seasonal variations. The numbers in Table 3 show that those classes can be easily identified by all the models. This is further supported by a visual inspection of Figure 5, illustrating how most of the object types are clearly detected, e.g. buildings with the roads connecting them. Figure 6 also shows that all compared models exhibit a certain level of uncertainty in classifying natural

Models	IoU [%]											
	bld.	pvs.	ips.	bs.	wt.	cfs.	dcs.	bsd.	vyd.	hvg.	agr.	pld.
U-Net	81.8	49.2	72.8	40.5	85.0	41.1	68.7	23.9	62.2	**48.4**	53.0	35.5
SegFormer	81.1	52.2	71.8	50.5	87.0	60.1	72.1	25.4	63.2	41.0	56.0	35.5
U-T&T	**81.9**	48.6	71.9	43.4	83.2	56.9	69.8	25.6	65.1	46.0	53.3	36.6
Swin	81.3	50.6	73.0	42.4	80.5	55.4	71.2	23.9	65.2	45.5	54.1	**38.9**
TSViT	78.7	48.7	68.8	51.5	85.2	62.4	69.7	21.5	64.2	41.0	55.8	36.3
T-ConvFormer (Ours)	81.0	**55.2**	**73.2**	**57.1**	**88.4**	**64.1**	**73.1**	**25.8**	**65.3**	42.2	**57.1**	37.1

Table 3. Class-wise IoU values [%] on the test set of the FLAIR #2 dataset achieved by different approaches. The compared methods are those defined in Table 2. The numbers are averages achieved by three independent test runs. Best results are indicated in **bold**.

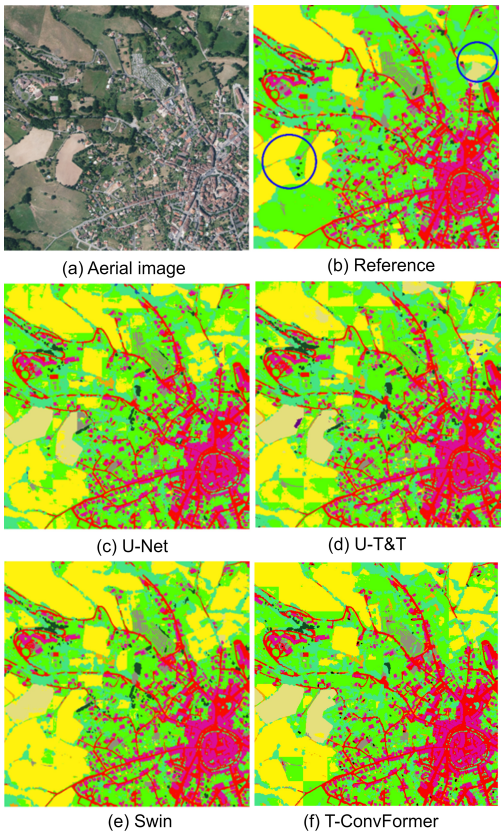

Figure 5. Aerial image of an urban test area, the corresponding reference and the land cover maps predicted by four selected methods. The area corresponds to multiple tiles that were classified independently. Blue circles show areas that are misclassified by all approaches. The acronyms for (c) – (f) correspond to the compared methods. Colours: magenta - *bld*, grey - *pvs*, red - *ips*, brown - *bs*, blue - *wt*, dark green - *cfs*, acquamarine - *dcs*, orange - *bsd*, purple - *vyd*, bright green - *hvg*, yellow - *agr*, dark yellow - *pld*.

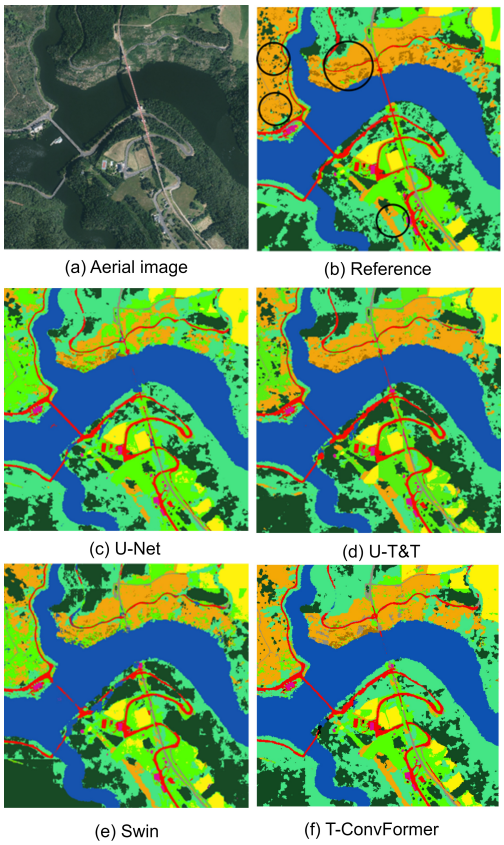

Figure 6. Aerial image of a rural test area, the corresponding reference and the land cover maps predicted by four selected methods. The area corresponds to multiple tiles that were classified independently. Black circles highlight regions that are classified better by approaches which integrate SITS. The acronyms for (c) – (f) correspond to the compared methods. Colour code: cf. Figure 5

areas such as *bs* and *bsd*, perhaps because they look similar to other object types, even when considering an entire vegetation cycle. Overall, despite remaining problems with certain classes or classification uncertainty near class boundaries, our results show the benefits of our method compared to methods from the state-of-the-art.

Table 4 presents the results of our ablation study, comparing our method to two variants as described in Table 1. The table shows that the integration of both components, cross-attention fusion and auxiliary supervision, improves the performance. We found that using a cross-attention approach in the fusion of SITS and aerial data improves the *mIoU* in the order of 0.8% compared to an element-wise addition approach (compare T-CF-1 to T-CF-0). Introducing intermediate auxiliary losses at all stages of the SITS branch decoder increases the model performance by another 1.0% in *mIoU* (T-CF-1 vs. T-ConvFormer). Similar improvements can be observed in the OA. In particular, the variant T-CF-0 not using the two components achieves a similar performance as the second best method according to Table 2, SegFormer, which does not use SITS. Our ablation study shows that the use of SITS in combination with our proposed new components leads to a significant performance improvement of about 1.7-1.8%.

6. CONCLUSION

In this paper, we presented a new method that combines convolutional and attention-based fusion networks to jointly use aer-

Name	mIoU [%]	OA [%]
T-CF-0	58.4 ± 0.0	72.5 ± 0.1
T-CF-1	59.2 ± 0.0	73.3 ± 0.2
T-ConvFormer	**60.1** ± 0.1	**74.3** ± 0.0

Table 4. Comparison of different variants of our approach. The variants are those defined in Table 1.

ial and SITS images for land cover classification. Overall, our results indicate that the integration of SITS improves the results, achieving an improvement of up to 1.9% in the *mIoU* compared to an approach only relying on aerial images and applying a SegFormer (Strudel et al., 2021). A comparison of the proposed approach with other transformer-based approaches from the literature shows a significant advantage of the former with respect to integrating the SITS and aerial data. The largest impact of the proposed model were remarkably seen on classes such as *bare soil*, *coniferous*, *vineyard*, *agricultural land* that particularly change over time. This improvement can be attributed to a combination of different components, including attention-based fusion and network supervision. This shows the benefit of combining multiscale data from multiple sensors as an efficient way to improve the classification of land cover.

The results presented in this work also indicate that, whereas SITS improve the classification accuracy significantly, the improvement is not very large in absolute terms, and some classes are relatively poorly differentiated. One of the reasons could be that the ratio between the GSD of the used data (a factor of 50 in this study) could be too large. Future work could look at combining aerial imagery with other higher resolution satellite images such as Planet Labs (Toker et al., 2022), which offer more fine-grained details about the land cover up to 3 m GSD. Another aspect to consider would be to investigate existing self-supervised learning approaches, with the advantage of using pre-trained models on large datasets. Beyond that, another topic to be addressed could be to improve the resolution of the SITS data using approaches for super-resolution (Okabayashi et al., 2024), e.g. based on diffusion models (Moser et al., 2024). Such methods could use the aerial data as additional input and might help to generate higher-resolution SITS images so that the gap between the spatial resolutions becomes smaller.

References

Benedetti, P., Ienco, D., Gaetano, R., Ose, K., Pensa, R. G., Dupuy, S., 2018. M3Fusion: A deep learning architecture for multiscale multimodal multitemporal satellite data fusion. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(12), 4939–4949.

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.

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. *International Conference on Learning Representations (ICLR)*.

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., Meygret, A., Spoto, F., Sy, O., Marchese, F., Bargellini, P., 2012. Sentinel-2: ESA's optical high-resolution mission for GMES operational services. *Remote Sensing of Environment*, 120, 25-36.

Garioud, A., Gonthier, N., Landrieu, L., De Wit, A., Valette, M., Poupée, M., Giordano, S. et al., 2024. Flair: A country-scale land cover semantic segmentation dataset from multi-source optical imagery. *Advances in Neural Information Processing Systems*, 36, 16456–16482.

Garioud, A., Peillet, S., Bookjans, E. M., Giordano, S., Wattrelos, B., 2022. FLAIR #1: Semantic segmentation and domain adaptation dataset. *ArXiv*, abs/2211.12979.

Garnot, V., Landrieu, L., Giordano, S., Chehata, N., 2020. Satellite image time series classification with pixel-set encoders and temporal self-attention. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 12322–12331.

Garnot, V. S. F., Landrieu, L., 2021. Panoptic segmentation of satellite image time series with convolutional temporal attention networks. *IEEE International Conference on Computer Vision (ICCV)*, 4852–4861.

Garnot, V. S. F., Landrieu, L., Chehata, N., 2022. Multi-modal temporal attention models for crop mapping from satellite time series. *ISPRS Journal of Photogrammetry and Remote Sensing*, 187, 294–305.

Gbodjo, Y. J. E., Montet, O., Ienco, D., Gaetano, R., Dupuy, S., 2021. Multisensor land cover classification with sparsely annotated data Based on Convolutional Neural Networks and self-distillation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14, 11485–11499.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *IEEE International Conference on Computer Vision (ICCV)*, 1026–1034.

Heidarianbaei, M., Kanyamahanga, H., Dorozynski, M., 2024. Temporal ViT-U-Net Tandem Model: Enhancing Multi-Sensor Land Cover Classification Through Transformer-Based Utilization of Satellite Image Time Series. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 10, 169–177.

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.

Ji, S., Zhang, C., Xu, A., Shi, Y., Duan, Y., 2018. 3D convolutional neural networks for crop classification with multi-temporal remote sensing images. *Remote Sensing*, 10(1), 75.

Kanyamahanga, H., Rottensteiner, F., 2024. Land Cover Classification based on Multiscale Time Series of Satellite and Aerial Images. *Proceedings, 44th Annual Scientific and Technical Conference of the DGPF in Remagen*, 32, 223–235.

Kingma, D. P., Ba, J., 2015. Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*.

Li, R., Zheng, S., Duan, C., Wang, L., Zhang, C., 2022. Land cover classification from remote sensing images based on multi-scale fully convolutional network. *Geo-spatial Information Science*, 25, 278–294.

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

MacDonald, E., Jacoby, D., Coady, Y., 2024. VistaFormer: Scalable Vision Transformers for Satellite Image Time Series Segmentation. *arXiv preprint arXiv:2409.08461*.

Moser, B. B., Shanbhag, A. S., Raue, F., Frolov, S., Palacio, S., Dengel, A., 2024. Diffusion models, image super-resolution, and everything: A survey. *Transactions on Neural Networks and Learning Systems (IEEE), early access*, 1-21. doi.org/10.1109/TNNLS.2024.3476671.

Okabayashi, A., Audebert, N., Donike, S., Pelletier, C., 2024. Cross-sensor super-resolution of irregularly sampled sentinel-2 time series. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 502–511.

Ren, B., Liu, B., Hou, B., Wang, Z., Yang, C., Jiao, L., 2024. SwinTFNet: Dual-stream transformer with cross attention fusion for land cover classification. *IEEE Geoscience and Remote Sensing Letters*, 21, 1-5.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Part III 18*, 234–241.

Rußwurm, M., Körner, M., 2017. Multi-temporal land cover classification with long short-term memory neural networks. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-1/W1, 551–558.

Sharma, A., Liu, X., Yang, X., 2018. Land cover classification from multi-temporal, multi-spectral remotely sensed imagery using patch-based recurrent neural networks. *Neural Networks*, 105, 346–355.

Strudel, R., Garcia, R., Laptev, I., Schmid, C., 2021. Segmenter: Transformer for semantic segmentation. *IEEE International Conference on Computer Vision (ICCV)*, 7262–7272.

Tarasiou, M., Chavez, E., Zafeiriou, S., 2023. ViTs for SITS: Vision transformers for satellite image time series. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 10418–10428.

Toker, A., Kondmann, L., Weber, M., Eisenberger, M., Camero, A., Hu, J., Hoderlein, A. P., Şenaras, Ç., Davis, T., Cremers, D. et al., 2022. Dynamicearthnet: Daily multi-spectral satellite dataset for semantic change segmentation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 21158–21167.

Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I., 2017. Attention is All you Need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.

Voelsen, M., Rottensteiner, F., Heipke, C., 2024. Transformer models for Land Cover Classification with Satellite Image Time Series. *PFG–Journal of Photogrammetry, Remote Sensing and Geoinformation Science*, 92(5), 547–568.

Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L., 2021. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *IEEE International Conference on Computer Vision (ICCV)*, 568–578.

Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J., 2018. Unified perceptual parsing for scene understanding. *European Conference on Computer Vision (ECCV)*, 418–434.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., Luo, P., 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34, 12077–12090.