Land Cover Classification based on Multiscale Time Series of Satellite and Aerial Images

HUBERT KANYAMAHANGA[1] & FRANZ ROTTENSTEINER[1]

Abstract: Recent advances in remote sensing technology have increased the availability of high-quality image data, which might have different characteristics and thus can provide complementary information for the same observed region. This paper presents methods for land cover classification based on the joint use of high-resolution aerial images and satellite image time series (SITS). We extend an existing approach by introducing transformer-based components, comparing two approaches that fuse features extracted from SITS and an aerial image before predicting land cover at the geometrical resolution of the aerial image. We perform experiments on an existing benchmark dataset, showing that the transformer-based fusion of an aerial image with a SITS from Sentinel-2 improves the classification results by +1.8% in the mean IoU and by +0.8% in the overall accuracy compared to fully convolutional networks based on aerial images only.

1 Introduction

Land cover classification, the task of assigning a class label representing the physical material of the Earth surface to each pixel in the image, is one of the most important tasks in remote sensing. With the growing availability of high-quality image data, multiple sensors can be used to acquire data with complementary information of the same observed region. For example, aerial imagery can deliver textural information at decimetre resolution, but usually with high revisit times. On the other hand, satellite systems have short revisit times, so that the resultant images can capture temporal changes and patterns, but usually at a coarser spatial resolution, e.g. with a ground sampling distance (GSD) of 10 m or more. Both aerial and satellite images can be combined to improve land cover classification results.

For aerial images, Fully Convolutional Networks (FCNs) with encoder-decoder architectures are frequently used for land cover classification, e.g. architectures based on U-Net (RONNEBERGER et al. 2015) that use skip connections between feature maps at corresponding resolutions from the encoder and the decoder to improve the spatial accuracy of the results. For SITS data, different methods such as 3D-Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs) have been used to extract spatial and temporal information. Methods based on 3D-CNNs consider the time as just an additional dimension of the input data and learn filter kernels for a convolution in the two spatial and the temporal dimensions (JI et al. 2018; LI et al. 2022). RNNs are designed to explicitly model sequential data and capture temporal dependencies of the time series while processing one image at a time, maintaining a memory of the previous inputs and updating the output based on the current input (ZHU et al. 2021). Recently, RNNs have been challenged by vision transformers (ViTs), which have been adapted to efficiently capture both spatial and temporal dependencies in SITS data, yielding promising results in several remote sensing tasks (VOELSEN et al., 2023; TARASIOU et al. 2023; ZHANG et al. 2023). However, the methods cited so far only use a single modality as input.

[1] Institut für Photogrammetrie und GeoInformation, Leibniz Universität Hannover, Nienburger Str. 1, D-30167 Hannover, E-Mail: [kanyamahanga, rottensteiner]@ipi.uni-hannover.de

The goal of this paper is to present a method for the simultaneous use of an aerial image and a co-registered Satellite Image Time Series (SITS) to obtain a land cover map at the GSD of the aerial image. Existing approaches for integrating such multi-scale images tend to use an architecture consisting of a separate network branch for each data modality (BENEDETTI et al. 2018; GBODJO et al. 2021; BERGAMASCO et al. 2023; GARIOUD et al. 2023) before fusing the outputs for land cover classification. Our method is based on (GARIOUD et al. 2023), which we also use as our baseline. The method uses a FCN branch for processing aerial images and another CNN-based branch for the SITS data before fusing the results for classification. The SITS branch uses an attention-based model for encoding temporal information. We extend (GARIOUD et al. 2023) by introducing transformer models for the SITS branch of the network, adopting the approaches from (VOELSEN et al. 2023) and (TARASIOU et al. 2023) to implement two different transformer-based architectures. The first one uses a modified Swin Transformer (VOELSEN et al. 2023), whereas our second architecture replaces the SITS branch of (GARIOUD et al. 2023) by the Vision Transformer for SITS (TSViT) of (TARASIOU et al. 2023). In both new approaches, the branch processing the aerial imagery and the fusion of the feature maps and the aerial images are identical to (GARIOUD et al. 2023).
The scientific contributions of this paper can be formulated as follows:

- We present two new methods for the integration of multi-scale data by jointly using aerial images and multi-temporal information from SITS to exploit their potential for improving land cover classification.
- In this context, we investigate two different transformer-based architectures, including the use of input time acquisitions of varying length for SITS.
- In experiments based on the French Land cover from Aerospace ImageRy (FLAIR) #2 Challenge dataset (GARIOUD et al. 2023) we analyse the impact of using SITS on the quality of the results of land cover classification, and we compare our two transformer-based architectures with each other and with our baseline (GARIOUD et al. 2023).

2 Related Work

We start this review with discussing related work that uses CNNs for integrating multiscale data from different sensors for the classification. We then introduce transformer models and the way in which they are adapted to the task of semantic segmentation for remote sensing images, in particular SITS. Finally, we discuss the few existing approaches that use attention-based modules to combine aerial and satellite image time series data.
The integration of multiscale remote sensing data sources has been investigated in several works. For instance, BENEDETTI et al. (2018) use a Gated Recurrent Unit (GRU) 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 network to obtain a land cover classification at the GSD of SPOT-6. GBODJO et al. (2021) combine Sentinel-2 and SPOT-6 data with SITS from Sentinel-1 images. The Sentinel-1 and SPOT data are processed by 2D-CNN based encoders, while the Sentinel-2 SITS is analyzed using a 1D-CNN encoder applying the convolution only in the temporal dimension. The resultant features are concatenated and used to predict land cover at the GSD of the SPOT images. Both approaches are limited to a fixed number of timesteps. Also, the difference in the GSDs is relatively small (10 m vs. 2 m GSD).

There are not many CNN-based approaches that combine SITS data with aerial images. BERGAMASCO et al. (2023) use a 3D-CNN network to extract spatial and temporal features from Sentinel-2 SITS, combining the features extracted from aerial images (0.2 m GSD) using a residual network (2D-ResNet). The fused features are fed to a decoder to discriminate different types of pasture at the GSD of the aerial images. The results show that combining aerial images and SITS data systematically improves the classification results compared to mono-modal classification. However, the method has limitations in differentiating classes that are semantically similar, and the class structure is not typical for land cover classification.
Transformer models utilize self-attention modules to model long-range dependencies and relationships within input sequences and have proven to outperform other models for processing and analysing sequential data (KHAN et al. 2022). Compared to 3D-CNNs, which typically require fixed input dimensions, they offer the advantage of being able to adapt to varying sequence lengths. Several approaches have attempted to use attention-based models for the extraction of spatial-temporal information from SITS. VOELSEN et al. (2023) extended Swin Transformer (LIU et al., 2021) for processing the 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 transformer 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 than other techniques, it has a quadratic complexity with respect to the input size, which can result in higher hardware demand when working with larger inputs. Neither VOELSEN et al. (2023) nor TARASIOU et al. (2023) combine multiple modalities at different GSDs.
Very few approaches have used attention-based models to jointly combine SITS data with aerial images. GARIOUD et al. (2023), on whose work ours is based, proposed a two-branch U-Net-based architecture to fuse Sentinel-2 SITS with aerial images. They adopt the U-TAE model of GARIOUD et al. (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 to the encoder features of all levels in the skip connections. We see a problem of (GARIOUD et al., 2023) in the way in which the attentions are computed; for all timesteps of a given SITS, the query (DOSOVITSKIY et al. 2021) is defined as the temporal average of the queries (all pixels) of all timesteps (GARNOT et al. 2019), which does not model the interaction between timesteps. To counteract this limitation, we propose to use transformer models (LIU et al. 2021; TARASIOU et al. 2023) in which the pixels of one timestep are encoded in a single query, which is then used to attend to all other elements of the timestep.
To the best of our knowledge, none of the existing approaches have investigated the use of transformer-based models to jointly integrate temporal information extracted from SITS data with features learned from aerial images for land cover classification. In this paper, we extend (GARIOUD et al. 2023) by introducing a fully attentional model for processing SITS based on transformer networks (LIU et al. 2021; TARASIOU et al. 2023). We compare two different variants of transformers and additionally evaluate the contribution of the satellite data to the classification results of aerial images.

3 Methodology

The main goal of our method is to combine SITS data and an aerial image for a pixel-wise prediction of land cover at the GSD of the aerial image. We propose a two-branch architecture which extends (GARIOUD et al. 2023) by introducing two different transformer-based approaches for processing SITS. This choice is motivated by the ability of transformer models to model temporal as well as long-range spatial dependencies.

In Section 3.1, we briefly describe our baseline model (GARIOUD et al. 2023), which also provides us with the general structure of the joint classification procedure. The two subsequent sections describe our two transformer-based models for the network branch for processing SITS. The first one, described in Section 3.2, replaces the SITS branch of (GARIOUD et al. 2023) by the transformer-based model of VOELSEN et al. (2023). The second one replaces that branch by the Vision Transformer for SITS (TSViT) of TARASIOU et al. (2023) and is presented in Section 3.3. Section 3.4 describes the training procedure used for all three approaches.

3.1 Joint Classification of Aerial Images and SITS using U-Net for SITS

The input of the method proposed in (GARIOUD et al., 2023) consists of a georeferenced SITS X^S with T timesteps, each image having C^S spectral bands and covering the same area of $H^S \times W^S$ pixels at the GSD of the SITS, and an aerial image X^A with C^A spectral bands and covering an area of $H^A \times W^A$ pixels at the resolution of the aerial image. The aerial image corresponds to a subset of the area covered by the SITS. The output of our method is a land cover map of dimension $H^A \times W^A$ at the GSD of the aerial image. An overview of the architecture is given in Figure 1. The input dimensions are set to $H^S = W^S = 40$ and $H^A = W^A = 512$ (cf. Section 4.1).

The SITS data are organized into a four-dimensional tensor of shape $T \times C^S \times H^S \times W^S$, which forms the input of the SITS branch. GARIOUD et al. (2023) use the U-Net-TAE architecture (GARNOT et al., 2021) for this branch, which is expected to extract temporal information from the given SITS. In this branch, each image of the time series is processed by a shared CNN encoder, and the lowest feature map is used to compute the temporal attention. The resulting attention masks are spatially upsampled; at each resolution, a weighted sum of the feature maps of the individual timesteps is computed, with the attentions being used as weights, and the resultant feature maps are concatenated to the feature maps of the corresponding layer in a CNN decoder. The final output of the decoder and, thus, the SITS branch is a feature map of dimension $64 \times H^S \times W^S$ at the spatial resolution of the SITS.

The aerial image is organized into a tensor of shape $C^A \times H^A \times W^A$ processed by a separate branch (aerial branch) which also has a U-Net structure with five resolution levels and skip connections between corresponding levels of the encoder and the decoder. The fusion of the features extracted from the SITS and those extracted from the aerial image is performed in the skip connections. For that purpose, the area of overlap between the SITS and the aerial image has to be cropped from the SITS feature map, and the cropped feature map is upsampled multiple times, so that there is one upsampled map for each resolution level of the U-Net of the aerial branch. The fusion itself consists of adding the upsampled SITS features to the output of the corresponding encoder level of the aerial branch (FM in Fig. 1). Thus, instead of just forwarding the encoder output to the corresponding layer in the decoder, the skip connection will additionally forward the multitemporal information encoded in the SITS features to the decoder. The features delivered by the last decoder layer are processed by a softmax layer to predict the class scores, on the basis of which the land use map is generated.

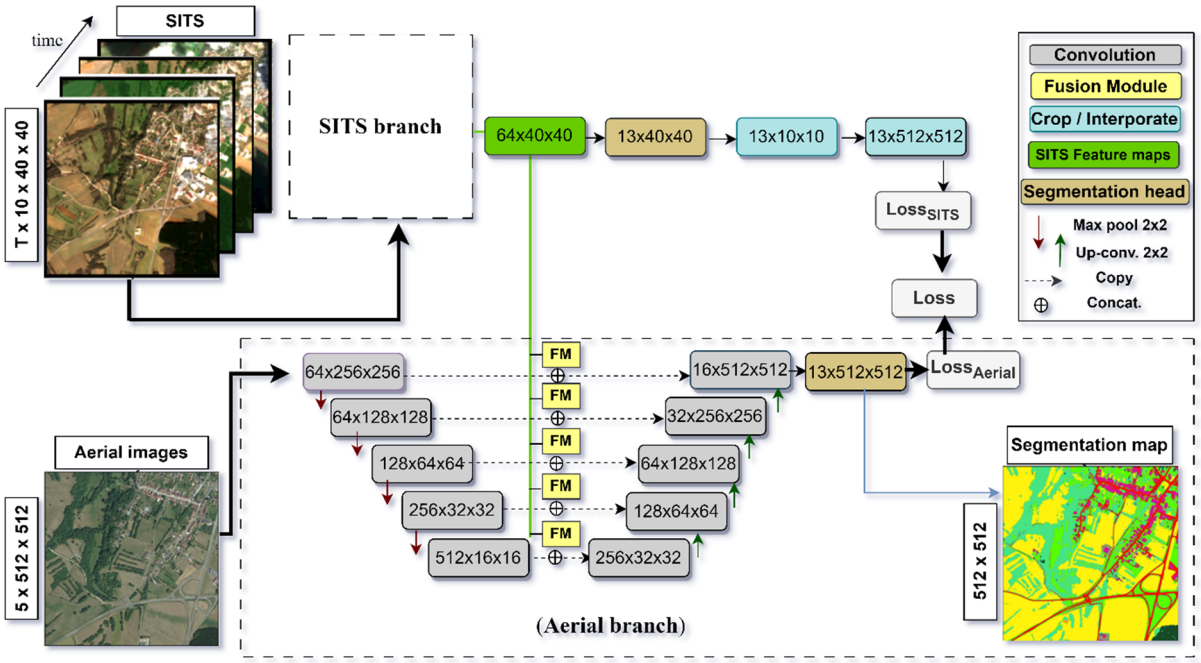


Fig. 1: Network architecture for the joint classification of SITS and an aerial image, adapted from (GARIOUD et al. 2023). The architecture of the SITS branch differs for the three approaches used in this paper. The aerial branch consists of a U-Net processing a single aerial image. This SITS feature map is upsampled to the resolutions of all the feature maps of the U-Net branch and added to the encoder output in the fusion modules (FM) situated in the skip connections. The combined features are used in the U-Net decoder, the output of which is used to produce the land cover map. In training, a classification loss is minimized for both the SITS and the aerial branches, but at test time, no labels are predicted for the SITS

3.2 SITS Branch based on a Swin Transformer

Our first method for the joint classification of an aerial image and a SITS uses the general architecture depicted in Figure 1, but it replaces the SITS branch of GARIOUD et al. (2023) by a transformer-based method. In particular, the SITS branch consists of an encoder based on a modified version of the Swin Transformer (LIU et al. 2021) and a decoder based on UPerNet (XIAO et al. 2018); the resultant architecture of the SITS branch is presented in Figure 2.

The original Swin Transformer computes attentions in a local window (e.g. patches of size 7x7). In order to still consider global context, consecutive attention layers are based on windows that are shifted by half the window (LIU et al. 2021). Our SITS branch is based on the extension proposed by VOELSEN et al. (2023). It consists of four processing stages, each generating a feature map at a different resolution (feature maps C1 … C4 with 1/4 … 1/32 of the size of an input patch). The temporal domain is only considered in the first stage, in which Swin Transformer blocks are executed in parallel to extract spatial features for each timestep and the outputs are fused and processed by a temporal transformer block introduced by VOELSEN et al. (2023). Stage 1 consists of two such blocks, considering shifted windows as in (LIU et al. 2021). The output is reshaped to a 3D tensor, discarding the temporal dimension, in the way described below. The result is a feature map C1 that is passed to the decoder before being downsampled to serve as an input for encoder stage 2. In the subsequent encoder stages, the standard Swin Transformer blocks are applied (two in stages 2 and 4, six in stage 3, following the tiny architecture of LIU et al (2021)). This combination corresponds to the network variant Swin_{S1} introduced in (VOELSEN et al. 2023), which outperformed other investigated variants in that publication.

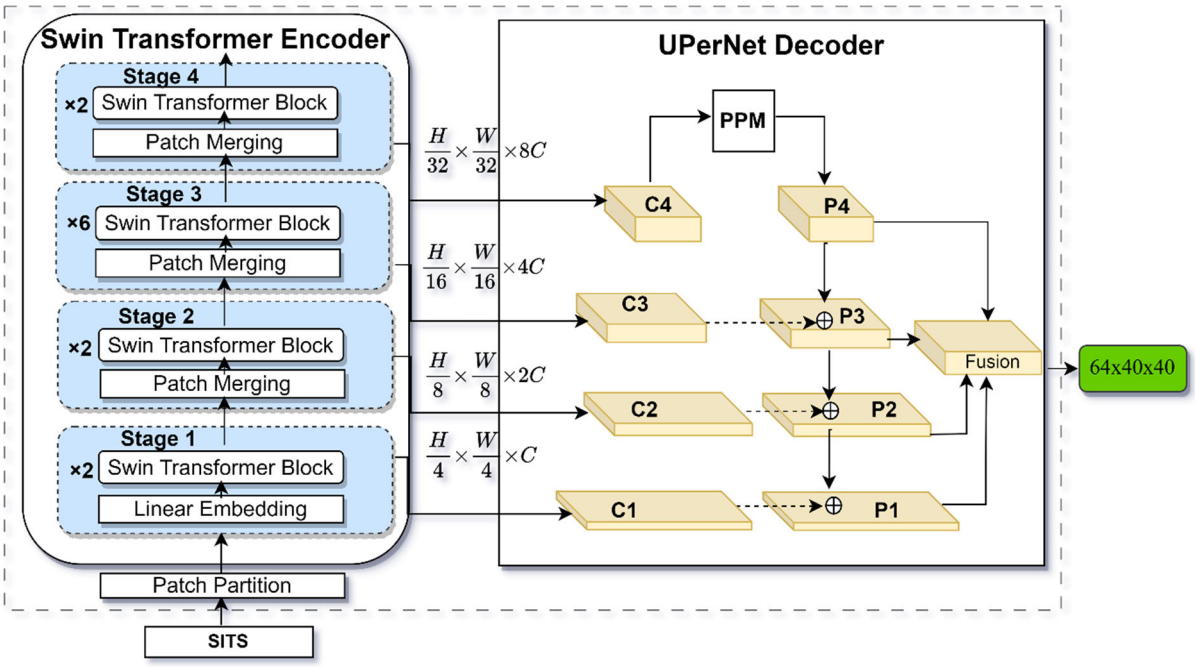


Fig. 2: Architecture of the SITS branch based on a Swin transformer encoder. It produces four feature maps C1, C2, C3, C4 that are used by a UPerNet decoder to generate a feature map of size (64 × 40 × 40) that is integrated into the Aerial branch.

The feature maps produced by the four stages of the encoder are used as input by the UPerNet decoder. The output of the last stage (C4) is processed by a Pyramid Pooling Module (PPM) (ZHAO et al. 2017) which applies average pooling using different window sizes and concatenates the results to the input after reducing the feature dimension by 1 × 1 convolutions. The resultant feature map P4 is upsampled to the resolutions of the feature maps C1, C2 and C3, and the upsampled version is concatendated with the corresponding feature map. The application of 1 × 1 convolutions results in three feature maps P1, P2, P3. These feature maps are upsampled to the resolution of P1 ($H^S/4 \times W^S/4$) and all of them (P1, P2, P3 and P4) are concatenated. Using a 1 × 1 convolution the number of features is reduced to 64, and the result is upsampled to the dimenison $H^S \times W^S$. This upsampled map is the output of the SITS branch. The main difference between the Swin_{S1} network of VOELSEN et al. (2023) and the Swin branch proposed here is that, we consider the temporal dimension only in the first stage; after that, the temporal dimension is collapsed. As a consequence of our approach, our architecture can deal with input sequences of varying length, i.e. T is variable, whereas in (VOELSEN et al., 2023) the input always has to consist of the same number of images (e.g. 4 or 12). The output of the last transformer block in stage 1 has a dimension of $T \times F^1 \times (H^S / 4 \cdot W^S / 4)$, where F^1 denotes the number of features. This is already three-dimensional, because the image patches were already arranged in a single dimension along the row direction. This tensor is reshaped to $F^1 \times (T \cdot H^S / 4 \cdot W^S / 4)$, and multi-head self-attention (VASWANI et al. 2017) is applied to generate another feature map of the same dimension. In this way, attention is considered between the feature vectors of all spatial positions. Afterwards, the feature maps corresponding to different timesteps are summed to merge the information from different timesteps. This results in a feature map of dimension $F^1 \times (H^S / 4 \cdot W^S / 4)$, which can be rearranged in any shape that is required by subsequent network layers.

3.3 Vision Transformer for SITS (TSViT)

Our second approach replaces the SITS branch of GARIOUD et al. (2023) by the Vision Transformer for SITS (TSViT) of TARASIOU et al. (2023). Each image of the SITS data is divided into non-overlapping patches, and attentions are computed between all timesteps of the corresponding patches at the same spatial location. Then, the outputs are reshaped and the attentions are computed between all patches of the same time step. We refer the reader to (TARASIOU et al. 2023) for more details about TSViT. For this approach, we also use a time series of varying length, similar to the Swin Transformer approach described in Section 3.2, and this is handled in the temporal encoder of TARASIOU et al. (2023). The branch processing the aerial imagery and the fusion of the feature maps from the SITS branch and the aerial images are identical to (GARIOUD et al., 2023).

3.4 Training

To train the proposed networks, a loss function L consisting of the sum of two terms, one for the SITS ($L_{CE,\ SITS}$) and one for the aerial network branch ($L_{CE,\ Aerial}$), is minimized. Both loss terms are based on a categorical Cross Entropy (CE) loss:

$$L_{CE,\ Branch} = \sum_{j=1}^{N_P} \sum_{i=1}^{N_C} t_{ij} \log p_{ij} \tag{1}$$

$$L = L_{CE,\ SITS} + L_{CE,\ Aerial}. \tag{2}$$

In equation (1), *Branch* can be *SITS* or *Aerial*, N_P is the number of pixels of an input patch of that branch, j is the index of a pixel, N_C is the number of classes, and i is the index of a specific class. The indicator variable t_{ij} indicates whether the reference class label of pixel j is i ($t_{ij} = 1$) or not ($t_{ij} = 0$), and p_{ij} is the softmax output for pixel j to correspond to class i. In order to be able to compute this loss, the SITS branch also has to predict class probabilities. Thus, a softmax layer is applied to the feature map predicted by that branch when the network is trained. As a reference is only available for the aerial image, which covers a smaller area than the SITS, the class scores are cropped to the area of overlap and then upsampled to the spatial resolution of the aerial image before computing the loss $L_{CE,\ SITS}$.

The loss in equation (1) is minimized using the Adam optimizer (KINGMA & BA 2015). For the SITS network, a random initialization strategy is used for its parameters while for the aerial network, the parameters are initialized starting from the weights pre-trained on the ImageNet dataset, similar to (GARIOUD et al. 2022).

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. 2023), 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 test areas distributed over 40 cantons in France, with a total area of about 817 km^2. All images 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 (BOA) 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 images per test area varies between 20 and 110. We follow the procedure used in our baseline method (GARIOUD et al. 2023), which requires a fixed-length input, and preprocess the SITS by computing monthly average reflections considering cloud-free pixels in the images, so that the maximum number of timesteps available for a test 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 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 *patches*) covering a size corresponding to 512 x 512 pixels in the aerial image. The SITS of every patch are sampled so that they have a size of 40 × 40 pixels at the GSD of 10 m. Altogether there are 77,762 patches, each with an aerial image, a SITS (with varying number T of images) and a reference label map. GARIOUD et al. (2023) defined a training set consisting of 61,712 patches and a test set consisting of the remaining 16,050 patches. We use the same definition, further splitting the 61,712 training patches into a set of 48,812 patches 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 patches. More details can be found in (GARIOUD et al. 2023).

4.2 Experimental Protocol

When applying the three methods described in Section 3 to the data described in Section 4.1, we used $H^A \times W^A = 512$ and $H^S \times W^S = 40$. The patch size for the tokens in the transformer-based methods is set to 2 x 2 pixels (TARASIOU et al. 2023). We use the split into training, validation and test sets as described earlier, which is also consistent with (GARIOUD et al. 2023). The training procedure was as described in Section 3.4. In training, we also applied data augmentation, using random rotations by 90^0, 180^0, 270^0, 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 30 epochs (early stopping). We used the ADAM optimizer (KINGMA & BA 2015) with the parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The batch size is set to 5 , the learning rate is set to 6e-5 and is decreased by a factor of 0.7 every 10 epochs. We have chosen those values because they were found to perform well on the validation dataset. The training is carried out on a cluster with 2x NVIDIA A100 80GB GPUs in a data parallel fashion where a distributed data parallel strategy is employed to leverage these computational resources efficiently, allowing for parallel training across multiple GPUs. All the models are implemented using the PyTorch Lightning Framework.

We carry out four sets of experiments. In the first one uses the U-Net of (GARIOUD et al. 2022) to predict land cover only based on the aerial images; it is referred to as *U-Net* in the remainder of this paper and its results will be compared against those of the other methods, all based on SITS, to assess the impact of the latter on the classification quality. The second set of experiments (referred to as *U-T&T*) is our baseline based on (GARIOUD et al. 2023). The set of

experiments referred to as *Swin* are based on the SITS branch described in Section 3.2, whereas the set *TSViT* uses method described in Section 3.3 for that purpose. 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 evaluate the performance of our models, the classification results on the test patches are compared to the reference. We report the intersection over union (IoU_C) for each class c:

$$IoU_C = \frac{TP_C}{TP_C + FP_C + FN_C} \quad (3)$$

In Equation 3, TP_C, FP_C, and FN_C denote the number of pixels that are true positives, false positives and the false negatives, respectively, for a class c. The mean intersection over union (*mIoU*) is also computed by taking the average of the IoU_C values for all classes except *other*, following the protocol in (GARIOUD et al. 2023). We also determine 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.

4.3 Results

Tab. 1 shows the *mIoU* and *OA* values achieved in the four experiments described above. The numbers show that the use of SITS data as an additional source of information leads to an increase in the overall performance. When using only aerial imagery in experiment *U-Net*, the lowest *mIoU* score of 55.2% is achieved. Using the baseline of GARIOUD et al. (2023) which also considers SITS (experiment *U-T&T*), the *mIoU* is slightly improved (56.8 % versus 55.2%), but it also has a higher standard deviation in the results. The methods applying transformer-based SITS branches perform on a similar level. The TSViT model achieved a slightly better *mIoU* score of 57.0 % than the one based on the Swin Transformer (56.9%), which also seems to be less stable, as indicated by the somewhat larger standard deviation. It could be said that when using the TSViT transformer, there is a significant improvement in *mIoU* (+1.8%) when using SITS compared to a model only relying on aerial imagery, whereas the differences between the methods that combine SITS and aerial data is not significant. As to be expected, the differences in *OA* are even smaller. The values seem to be more stable across different test runs, as indicated by the standard deviations. Again, the numbers indicate that the SITS help (improvement of +0.6-0.8% for the two transformer-based methods), but the differences between the methods that combine SITS and aerial images are barely significant.

Tab. 1: Mean *IoU* (*mIoU*) and Overall Accuracy (*OA*) for land cover classification [%] on the test set of the FLAIR #2 dataset achieved by different deep learning methods. The numbers are the averages achieved by three independently trained models, followed by the standard deviation. *U-Net*: the U-Net model using only aerial images. *U-T&T*: the baseline (GARIOUD et al. 2023). *Swin*: Our model based on the Swin Transformer (cf. Section 3.2). *TSViT*: our model using the TSViT transformer. All except *U-Net* use both aerial and satellite imagery. Best results are indicated in bold.

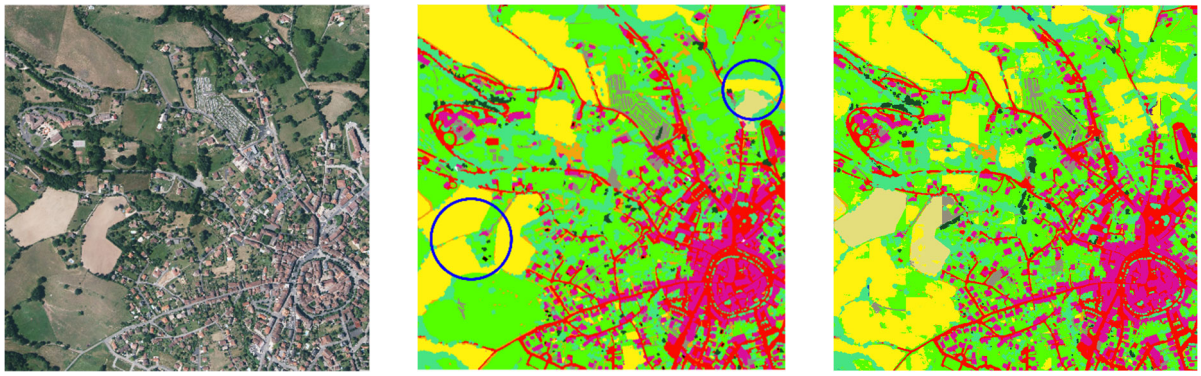
Method	*mIoU* [%]	*OA* [%]
U-Net	55.2±0.0	71.3±0.0
U-T&T	56.8±0.7	71.7±0.0
Swin	56.9±1.1	**72.1**±0.1
TSViT	**57.0**±0.0	71.9±0.0

Tab. 2 presents the class-wise *IoU* values achieved in the four experiments, averaged over the three test runs. The numbers in the table indicate that there is no unique tendency in the

performance of the compared methods w.r.t. the different classes. Nevertheless, the method solely based on aerial images only achieves the best IoU value for class *hvg*. The method based on the Swin Transformer preforms best for five classes, the one based on TSViT for four and the baseline for two. In general, the improvements due to the consideration of SITS is in the order of 1%-2% across classes, but it can be up to 20% for *cfs*. To a certain degree this can be expected, because coniferous trees show a different seasonal behaviour than some other vegetation classes, which would not be reflected in the aerial data. The fact that the scores for the majority of the classes are better for the transformer-based methods would indicate a slight advantage of these approaches over the baseline, but as pointed out earlier, on average the differences are quite small. Fig. 3 presents qualitative results for an urban area. An example for a rural area is shown in Fig. 4.

Tab. 2: Class-wise IoU values [%] on the test set of the FLAIR #2 dataset achieved by different deep learning methods. The compared methods are those defined in Tab. 2. The numbers are averages achieved by three independently trained models. Best results are indicated in bold.

Method	*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
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

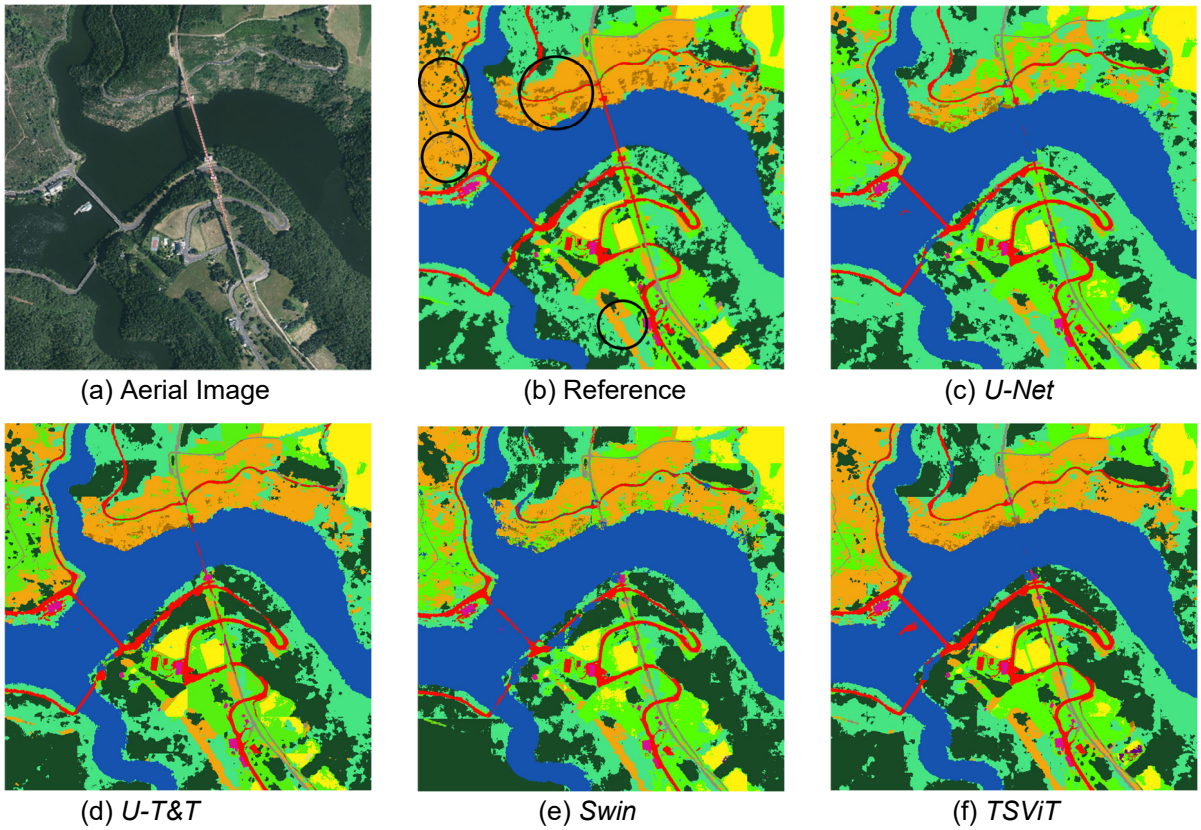


(a) Aerial Image (b) Reference (c) *U-Net*

(d) *U-T&T* (e) *Swin* (f) *TSViT*

Fig. 3: Aerial image of a test area, the corresponding reference and the land cover maps predicted by the four methods compared in our experiments. The area corresponds to multiple patches that were classified independently. It is dominated by land cover that is typical for urban areas. 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*

Tab. 2 also shows that some classes can be differentiated more easily than others. The frequency of the classes in the data seems to matter, as some of the highest *IoU* scores (55%-81%) are achieved for classes occurring very frequently (e.g., *ips*, *dcs*, *bld*, and *agr*), whereas some of the lowest scores are achieved by underrepresented classes (e.g. *bsd*, *pld*). Nevertheless, *hvg*, a class with a high frequency of occurrence, achieves a very low *IoU* score, whereas *vyd*, an underrepresented one, achieves a relatively high one. This may be attributed to a very similar appearance of some classes. For example, it might be difficult to differentiate herbaceous vegetation areas (e.g. gardens, public parks) against agricultural land in some areas.



(a) Aerial Image (b) Reference (c) *U-Net*

(d) *U-T&T* (e) *Swin* (f) *TSViT*

Fig. 4: Aerial image of a test area, the corresponding reference and the land cover maps predicted by the four methods compared in our experiments. The area corresponds to multiple patches that were classified independently. It is dominated by land cover that is typical for rural areas. Black circles highlight regions that are classified better by approaches which integrate SITS. The acronyms for (c) – (f) correspond to the compared methods. The colour code is identical to the one in Fig. 3

Fig. 3 and 4 show how the use of temporal information from SITS data helped in identifying classes which change over time. It can be seen that *agr*, *vyd*, *cfs*, are better separated by approaches that integrate SITS data (cf. the regions in the black circles). In general, all the models show similar performance on classes such as *bld*, *pvs*, *ipv*, and *bsd*, which are not affected by seasonal changes. The numbers in Table 2 show that these classes can be easily identified by all the models This is also confirmed by a visual inspection of Figure 3, showing how most of the object types are clearly detected, e.g. buildings with the roads connecting them. Fig. 4 also show that all models exhibit a certain level of uncertainty to classify natural areas such as *bs* and *bsd*, potentially because they look similar to other object types, even when considering an entire vegetation cycle. In summary, despite some remaining problems, our results show the benefits of integrating SITS for land cover classification using transformers.

5 Conclusion and Outlook

In this paper, we investigated two different approaches based on transformer models to integrate aerial and SITS for land cover classification. Our results show that the integration of SITS improves the results by a margin of up to 1.8% in *mIoU* compared to what can be achieved by an approach only relying on aerial images and applying a classical U-Net (GARIOUD et al. 2022). A comparison of the two transformer-based models and the baseline (GARIOUD et al. 2023) shows a slight advantage for the former, though the overall improvement is relatively small. The largest improvement due to the integration of SITS was achieved for *cfs*, but classes such as *bs*, *vyd* or *agr* are also classified. This indicates the benefit of combining the two modalities for the classification of land cover.

Our results also indicate that, whereas SITS improve the classification accuracy that can be achieved, this improvement is relatively small, and some classes are relatively poorly differentiated. Future work could investigate different ways of computing attentions for extracting temporal information from SITS data. Also, the way in which the timesteps are combined after stage 1 of the Swin Transformer encoder could be changed to allow a multi-temporal information flow at higher levels of the network. Another issue to be addressed could be the way in which the features of the SITS and the aerial data are fused; perhaps, attention-based models could also be used here. Finally, the training procedure could be changed, e.g. by using alternative loss functions that compensate for the class imbalance of the training data.

6 References

BENEDETTI, P., IENCO, D., GAETANO R., OSE K., PENSA R. G. & DUPY, 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., ZHAII, 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. ArXiv:2010.11929.

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. Rem. Sen. of Environment, **120**, 25-36.

GARIOUD, A., PEILLET, S., BOOKJANS, E., GIORDANO, S. & WATTRELOS, B., 2022: FLAIR #1: Semantic Segmentation and Domain Adaptation. ArXiv, abs/2211.12979.

GARIOUD, A., PEILLET, S., BOOKJANS, E., GIORDANO, S. & WATTRELOS, B., 2023: FLAIR #2: Textural and Temporal Information for Semantic Segmentation from Multi-source Optical Imagery. ArXiv preprint 2305.14467.

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), 4872-4881.

GARNOT, L. LANDRIEU, 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.

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.

JI, S., ZHANG, C., XU, A., SHI, Y. & DUAN, Y., 2018: 3D Convolutional Neural Networks for Crop Classification with MultiTemporal Remote Sensing Images. Remote. Sens., **10**, 75.

KHAN, S., NASEER, M., HAYAT, M., ZAMIR, S. W., KHAN, F. S. & SHAH, M., 2022: Transformers in Vision: A Survey. ACM Computing Surveys, **54**(10s), 1-41.

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**(2), 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.

RONNEBERGER, O., FISCHER, P. & BROX, T., 2015: U-Net: Convolutional Networks for Biomedical Image Segmentation. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015, **III**, 234-241.

STRUDEL, R., GARCIA, R., LAPTEV, I. & SCHMID, C., 2021: Segmenter: Transformer for Semantic Segmentation. IEEE International Conference on Computer Vision, 7262-7272.

TARASIROU, 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.

VASWANI, A., SHAZEER, N. M., PARMAR, N., USZKOREIT, J., JONES, L., GOMEZ, A. N., KAISER, L. & POLOSUKHIN, I., 2017: Attention is All You Need. 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 Ann. Photogramm. Remote Sens. Spatial Inf. Sci., **X**-1-W1-2023, 981-990.

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.

ZHANG, F., WANG, Y., DU, Y. & ZHU, Y., 2023: A Spatio-Temporal Encoding Neural Network for Semantic Segmentation of Satellite Image Time Series. Applied Sciences, **13**(23), 12658.

ZHAO, H., SHI, J., QI, X., WANG, X. & JIA, J., 2017: Pyramid Scene Parsing Network. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 6230-6239.

ZHU, Y., GEIß, C., SO, E., JIN, Y., 2021: Multitemporal Relearning with Convolutional LSTM Models for Land Use Classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, **14**, 3251-3265.