MultiGFM: multi-temporal framework for multi-modal geospatial foundation models

REIKO LETTMODEN[1], PEDRO ACHANCCARAY[1], KSENIA BITTNER[2] & MARKUS GERKE[1]

Abstract: Multi-temporal multi-modal Geospatial Foundation Models (GFMs) leverage data from different sources and satellite image time series (SITS) but face challenges such as extracting complex features from high-resolution modalities or being too computationally expensive. Mono-temporal GFMs address these issues using a single spatial encoder across multiple modalities to reduce computational overhead. To enhance this, we propose MultiGFM, a framework that introduces SITS into mono-temporal multi-modal GFM. Our experiments obtained the best results using Prithvi's spatial decoder and Lightweight Temporal Attention for temporal encoding. Using DOFA (Dynamic One-For-All) as the GFM for land use/land cover segmentation, we achieved a mIoU of 82.4% and gains of 1.8%/5.6% over the mono-temporal mono-/multi-modal models.

1 Introduction

Climate change and its associated extreme weather conditions have increasingly affected vegetation and ecosystems. For instance, in Europe, climate change causes changes in precipitation patterns, resulting in more extreme intensities and flooding or heatwaves with severe droughts (OVERBECK & SCHMIDT 2012; UFZ 2023; EEA 2022). In this context, remote sensing (RS) provides critical data for the creation of land use and land cover maps (LULC) for monitoring different classes, such as water bodies, trees, and croplands. For this purpose, RS data are acquired at different epochs in time (multi-temporal), from different sensors (multi-modal), such as passive (optical) and active (radar) sensors, and at different spatial resolutions (Sentinel-1: 10m, PlanetScope: 3m) to produce robust LULC maps (TSENG et al. 2023; GARNOT et al. 2022; GARIOUD et al. 2023). Finally, LULC maps are generated by deep learning (DL) methods trained in a supervised manner using datasets comprising numerous labeled RS data, where the annotation process is time-consuming, costly, and requires expert knowledge. Self-supervised learning (SSL) techniques have been employed in recent years to alleviate the need for large amounts of labeled RS data to train DL models. SSL allows models to learn from vast amounts of unlabeled data by creating pseudo-labels from the data itself through carefully designed pretext tasks. Some examples of pretext tasks are: contrastive learning (learn similar representations for different views of the same image) (BERG et al. 2024), masked image modeling (reconstruct missing parts of an image) (LI et al. 2024), and jigsaw puzzle solving (learn to arrange scrambled image patches) (ZHAO et al. 2022). Subsequently, once the model is trained, it is fine-tuned for a specific downstream task using a much smaller labeled dataset. Many geospatial foundation models (GFM) have adopted SSL techniques for pre-training on multi-modal and/or multi-temporal RS datasets for different downstream tasks, such as image classification, semantic segmentation, and object detection.

[1] Technische Universität Braunschweig, Institut für Geodäsie und Photogrammetrie, Bienroder Weg 81, D-38106 Braunschweig, Deutschland, E-Mail: r.lettmoden@posteo.de, [p.diaz, m.gerke]@tu-braunschweig.de

[2] Deutsches Zentrum für Luft- und Raumfahrt, Institut für Methodik der Fernerkundung, Oberpfaffenhofen, D-82234 Weßling, Deutschland, E-Mail: ksenia.bittner@dlr.de

Multi-temporal multi-modal GFMs leverage information from different sources and satellite image time series (SITS) but still face challenges such as dealing with different spatial resolutions per modality (GARNOT et al. 2022), extracting complex features from high-resolution modalities (e.g., OmniSat (ASTRUC et al. 2024)) or being too computationally expensive for consumer hardware (e.g., SkySense (GUO et al 2024)). On the other hand, mono-temporal multi-modal GFMs employ a single spatial encoder across multiple modalities, based on each modality embedding (e.g., msGFM (HAN et al. 2024)) or wavelength (e.g., DOFA (XIONG et al. 2024)), to reduce the computation overhead but lack learning temporal patterns associated with classes that change over time, such as croplands and trees due to their development over the year (phenological stages).
To further enhance multi-modal GFMs, we propose MultiGFM, a framework that introduces SITS into a mono-temporal multi-modal GFM. The proposed framework employs a GFM as a shared spatial encoder for each modality for robust feature extraction and a temporal encoder to capture temporal dependencies. We conducted several experiments to assess different spatial decoders, temporal encoders, data fusion techniques, and the importance of the different modalities in our framework. As a case study, we used the proposed framework for semantic segmentation to produce LULC maps in the Harz region, with a focus on tree monitoring.
The key contributions of this work can be summarised in the following three points:

- A novel framework to enhance multi-modal mono-temporal GFM by introducing SITS to exploit temporal dependencies further.
- Extensive assessment of the proposed framework by investigating the impact of different spatial decoders, fusion methods, modalities, temporal encoders, and their computational performance.
- Conducting a case study in the Harz region by producing LULC maps to monitor tree species and dead trees.

The remainder of this paper is organised as follows: Section 2 describes the dataset used in our experiments; Section 3 provides a detailed description of the proposed framework; Section 4 describes the experimental setup and summarizes the results obtained, where quantitative and qualitative analyses are performed; and Section 5 concludes the paper with a discussion of the results and our insights.

2 Dataset

The dataset used in our experiment comprised a SITS from the Harz Mountains and their surroundings. Fig. 1 shows the location of our study area in Germany (left) and a Sentinel-2 image (right) from the same area. The Harz Mountains are located at altitudes ranging from 200 to 640 m.s.l., with a mean annual temperature of 8°C, humid weather, and a mean annual precipitation of 760 mm (PUTZENLECHNER et al. 2024).
SITS data from different sensors covering the study area were acquired for 2020 and 2021. Tab. 1 summarizes the three sensors used (Sentinel-1 (S1), Sentinel-2 (S2), and PlanetScope) with different modalities and spatial resolutions. For optical images, only those with cloud cover lower than 5% were selected. For S2, Level-2A products were used with 10 spectral bands resampled to a 10m spatial resolution using the nearest neighbor method. For S1, we acquired dual-polarization (VV and VH) C-band Ground Range Detected images with a spatial resolution of 10m and Interferometric Wide Swath mode.

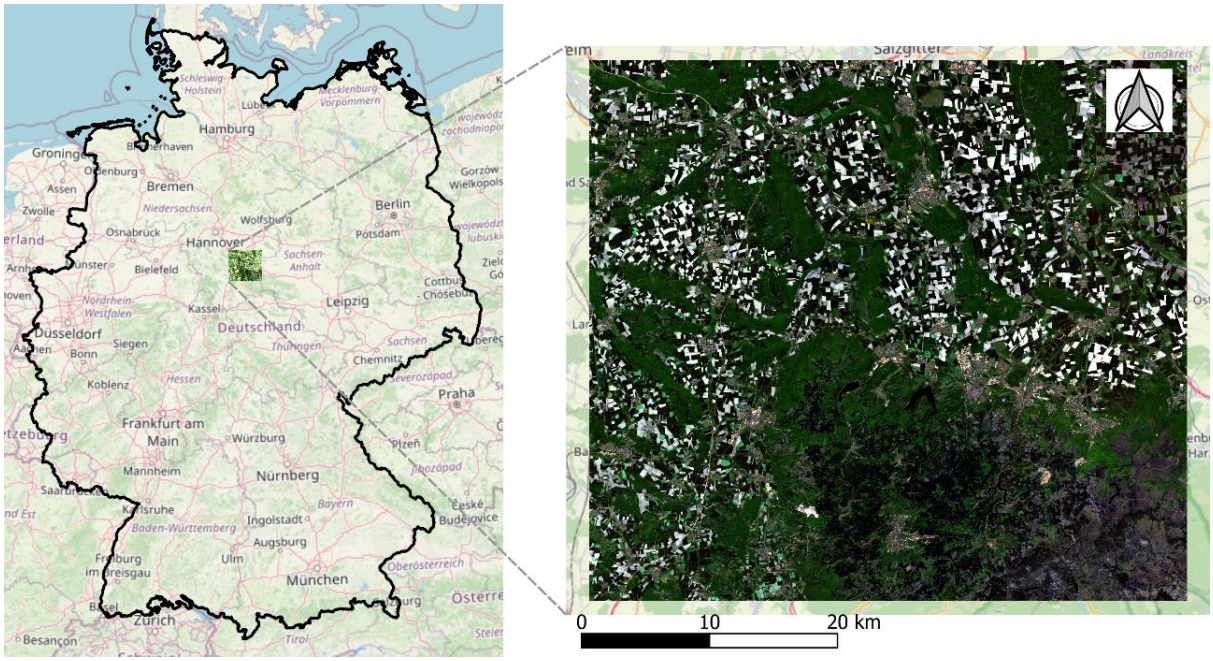

Fig. 1: Study area located in the Harz region, Germany (left), and a Sentinel-2 image of the same area.

Tab. 1: Information about the dataset used with different modalities and spatial resolutions.

Sensor	Modality	Spatial resolution (m)	# Images (2020-2021)	Bands
PlanetScope	Optical	3	7	R, G, B, NIR
Sentinel-2 (S2)	Optical	10	7	B2-B8a, B11, B12
Sentinel-1 (S1)	Radar	10	51	VV and VH

Each S1 image was pre-processed in the following manner: thermal noise removal, radiometric calibration, terrain correction, and a median filter with a window size 3×3 to reduce the speckle effect. We focused on growing season images (June – August) to reduce the effect of different weather conditions (cloud cover and snow) that could occlude the image and avoid confusion between a dead tree and a tree in a phenological stage where all its leaves fall.

The LULC classes considered in our study area were grass, crop, built-up, water, trees, deciduous trees, coniferous trees, and dead trees. Fig. 2 shows the images containing the labels for 2020 and 2021 with the aforementioned classes. For the non-tree-related classes, we used LULC maps from the European Space Agency (ESA) (ESA 2020; ESA 2021) with 10m spatial resolution. For trees, deciduous trees, and coniferous trees, we utilized the dominant tree species map from 2018 generated by BLICKENSDÖRDER et al. (2024), assuming that there were no significant changes from 2018 to 2020. Available tree species were classified into two main categories: coniferous and deciduous. For the dead tree class, we manually delineated dead tree samples using PlanetScope images with a 3m spatial resolution. Finally, the three sources (ESA LULC map, dominant tree species, and manual annotation) used for different classes were merged to produce a LULC reference map to train our proposed framework in a supervised manner. The class distribution, in percentage, in our dataset was the following: grass (12%), crop (38%), built-up (3.5%), water (1%), trees (7%), deciduous trees (18%), coniferous trees (18.5%), and dead trees (2%), where we can see it is highly imbalanced, especially for water and dead trees classes.

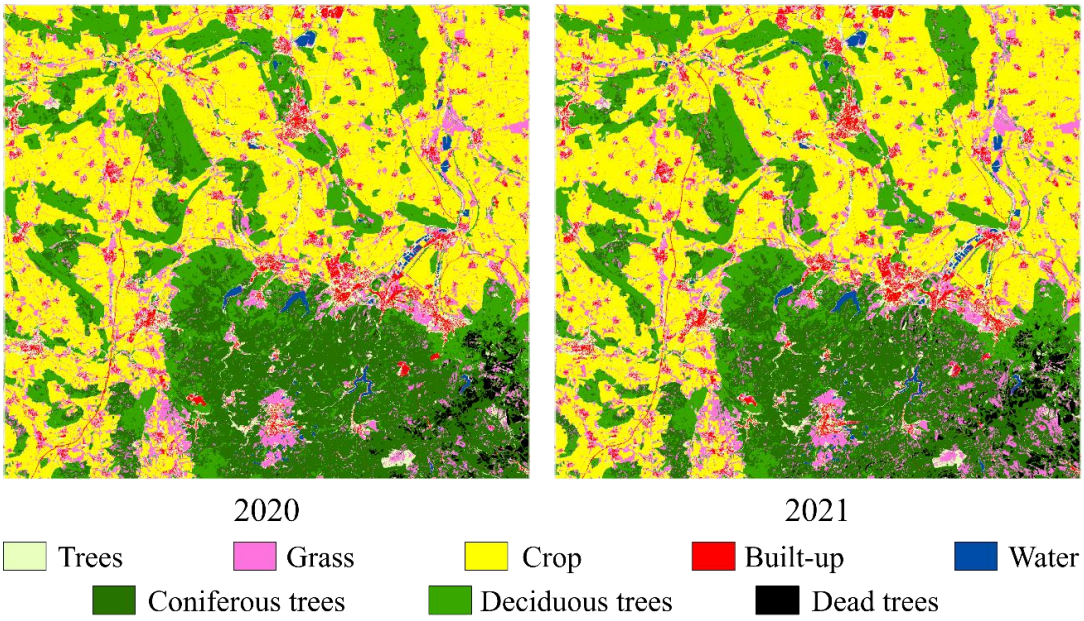

Fig. 2: Labels for the years 2020 and 2021.

3 Methodology

The methodology used in this study is shown in Fig. 3. The dataset comprises a SITS from different modalities (sensors), such as radar and optical, with different spatial resolutions. Each image from the dataset was pre-processed by extracting non-overlapping square patches. The entire geographical area covered by the images was divided into three non-overlapping sets: training, validation, and testing sets. All image patches from the same geographical location were assigned to the same set, regardless of the time they were captured. This ensured that the three sets were disjointed both spatially and temporally. From the training set, the mean and standard deviation were computed to apply standard normalization to the three sets. Subsequently, the training and validation sets were employed by the segmentation model to modify its weights and tune its hyperparameters, respectively. Finally, the model predictions are assessed quantitatively (based on the mean Intersection over Union – mIoU and F1-score) and qualitatively (visual assessment).

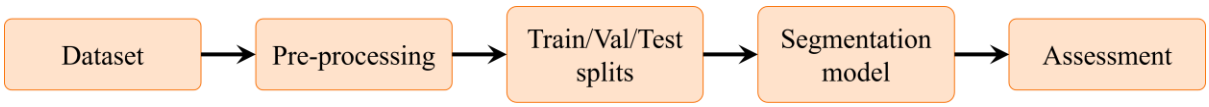

Fig. 3: Methodology followed in this work.

For the segmentation model, we propose a framework called MultiGFM to introduce RS SITS to mono-temporal multi-modal GFMs. The proposed framework is illustrated in Fig. 4. First, a shared spatial encoder (a multi-modal GFM) extracts spatial features from all patches from different modalities. The encoded spatial features are then decoded for the semantic segmentation task using different spatial decoders per modality. Later, the temporal dimension is collapsed using temporal encoders per modality. Finally, the spatial-temporal features are resized to the highest spatial resolution among all modalities and fused (concatenated) to be processed by the segmentation head, which provides a probability of belonging to each class for all pixels in the image.

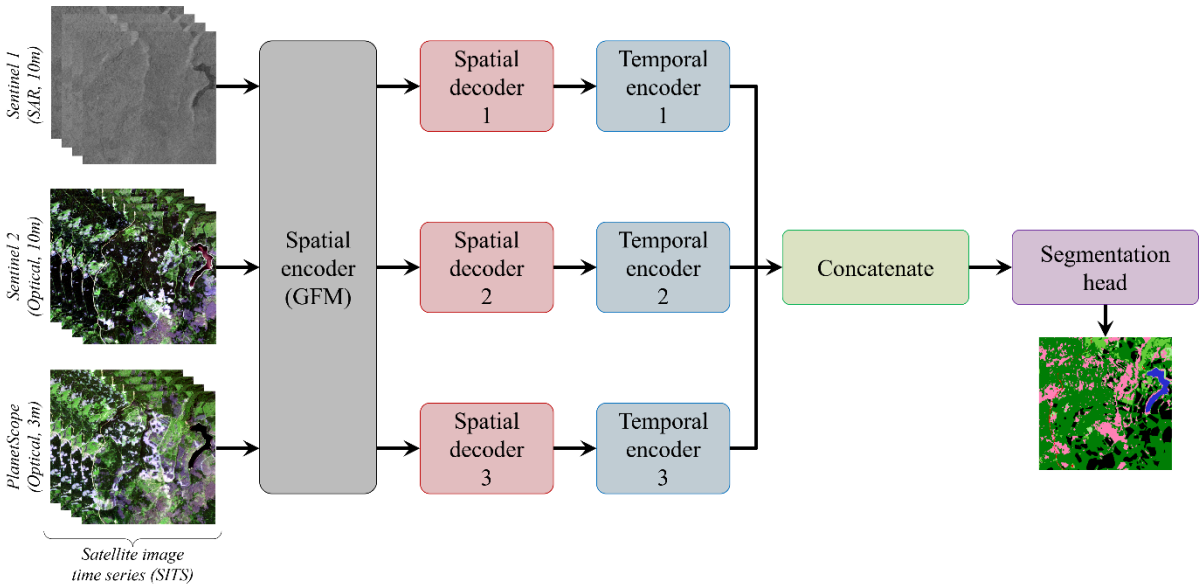

Fig. 4: Overview of the proposed framework, MultiGFM. Each SITS from different modalities is processed using a shared spatial encoder (a pre-trained multi-modal GFM). Then, the encoded spatial features are decoded for the semantic segmentation task using specific spatial decoders per modality, and the temporal dimension is collapsed using temporal encoders per modality. Finally, the spatial-temporal features of each modality are resized to the highest spatial resolution and then concatenated to be processed by the segmentation head.

4 Results

The proposed framework was employed for semantic segmentation, where different hyperparameters, data augmentation techniques, spatial decoders, temporal encoders, data fusion techniques, and effects of different modalities were assessed. All the experiments were conducted on a single node of the Phoenix cluster from TU Braunschweig. Each node contained 4 NVIDIA Tesla P100 16GB HBM2, 64GB RAM, and 2 CPU Intel Xeon E5-2640v4. The proposed framework was implemented using the GeoSeg framework based on PyTorch Lightning. Tab. 2 summarizes all the frameworks and models used and their corresponding repositories.

Tab. 2: Models, components, and frameworks used in our experiments with their corresponding GitHub repositories.

Model/Component/Framework	GitHub repository
DOFA (XIONG et al. 2024)	https://github.com/zhu-xlab/DOFA
UPerNet (XIAO et al. 2018)	https://github.com/yassouali/pytorch-segmentation
Prithvi (JAKUBIK et al. 2023)	https://github.com/NASA-IMPACT/hls-foundation-os
SegFormer (XIE et al. 2021)	https://github.com/open-mmlab/mmsegmentation
L-TAE (GARNOT & LANDRIEU. 2020)	https://src.koda.cnrs.fr/iris.dumeur/ssl_ubarn
fvcore: FLOPs calculation	https://github.com/facebookresearch/fvcore
GeoSeg	https://github.com/WangLibo1995/GeoSeg
PyTorch Lightning	https://lightning.ai/docs/pytorch/stable/

4.1 Experimental Setup

We used the Dynamic One-For-All (DOFA) model (XIONG et al. 2024) as a multi-modal GFM. The DOFA's spatial encoder (an enhanced ViT-B (DOSOVITSKIY et al. 2021)) and decoder

(UPerNet) are employed as spatial shared encoders and decoders. Late fusion (concatenation of spatial-temporal features) was used as the default data fusion method.
All models were trained for 100 epochs with a cosine annealing scheduler for the learning rate, model checkpoints based on the best validation mIoU, and cross-entropy loss as the loss function, with a label-smoothing factor of 0.1. As the optimizer, we used AdamW (LOSHCHILOV & HUTTER 2017) with a weight decay equal to 0.01 and a learning rate of $6e-5$ and $6e-6$ for the DOFA's wavelength encoder and spatial encoder, respectively. The batch size was set to 1 per GPU using the distributed data parallel (DDP) mode in a node with 4 GPUs, so the final batch size was equal to 4. To avoid the degeneration of batch normalization owing to a batch size of 1, synchronized batch normalization was used among all GPUs.
For data augmentation, flips and crop & resize (crop size equal to 0.75) transformations were applied to all modalities in the same manner, independent of their spatial resolution. Temporal dropout was applied as a regularization technique, similar to GARNOT et al. (2022), with $p_{optical} = 0.4$ and $p_{SAR} = 0.2$, where it was guaranteed that at least one patch per modality exists.
FLOPs were calculated using the fvcore library and taking one sample from 2021, as there are longer sequences in this year (30 S1, 4 S2, and 4 PlanetScope).
For S1 and S2, all the images were split into 128×128 patches. For PlanetScope, owing to the different spatial resolutions compared to S1 and S2, 384×384 patches were extracted. Training, validation, and test sets were created with ratios of 60%, 20%, and 20%, respectively. Patches containing non-data were discarded during pre-processing.

4.2 Experimental Protocol

The experimental protocol was designed to extensively assess the proposed framework and its main components: the spatial decoder, temporal encoder, hyperparameters, regularization, influence of each modality, weights initialization, and data fusion techniques. Furthermore, the number of FLOPs for each component was computed and included in the analysis to find a trade-off between a lightweight framework and robust performance in the test set.
For hyperparameter tuning, we compared two approaches: DOFA's default configuration (lower learning rate for the backbone ($lr_{backbone}$) than the entire model's learning rate (lr_{model})), and using the same learning rate (backbone and entire model trained with the same learning rate, which is the lr_{model}).
For regularization techniques, we compared three models: the baseline model, a model with data augmentation, and a model with data augmentation and temporal dropout.
For the spatial decoder, DOFA used a neck and UPerNet (XIAO et al. 2018). We compared DOFA's spatial decoder with decoders from SegFormer (XIE et al. 2021) and Prithvi (JAKUBIK et al. 2023), which are lightweight decoders. Prithvi's decoder is based on a series of transpose convolutions, where we used only the first blocks, whereas SegFormer's decoder is based on a Multi-Layer Perceptron (MLP) with linear and up-sampling layers.
Three techniques were employed for multi-modal data fusion: early, mid, and late fusion. For early fusion, we concatenated the closest sample in time for each modality to each sample of the modality with the shortest sequence length, where all samples were resized to the highest spatial resolution (PlanetScope, 3m, 384×384 patches). For mid fusion, similar to GARNOT et al. (2022), the spatial features provided by the shared spatial encoder per modality are concatenated following a time series. For late fusion, each modality is processed by a shared

spatial encoder, spatial decoders per modality, and temporal encoders per modality. The resulting features were then added (Late: Sum) or concatenated (Late: Concat). In addition, to measure the effect of using only a few S1 images, a configuration with a high S1 dropout ($p_{S1} = 0.9$) was included.

To measure the effect of each modality, all subsets of modalities were compared in terms of mIoU and computational efficiency (FLOPs). For temporal encoders, temporal max-pooling and L-TAE (GARNOT & LANDRIEU 2020) were used. These methods were compared to single temporal (ST) configurations using each modality and all modalities. For ST configurations, because the model does not extract any temporal features, the labels are repeated T times for a time series of length T.

For weights initialization, the model was trained with the ViT-B (DOSOVITSKIY et al. 2021) spatial encoder initialized with no pre-trained weights (trained from scratch), with pre-trained weights on ImageNet (DENG et al. 2009), and with the entire DOFA spatial encoder pre-trained as a foundation model. In this manner, we evaluated the advantages of using a pre-trained foundation model as a shared spatial encoder.

4.3 Quantitative Results

In the following subsections, we present our results in accordance with the aforementioned experimental protocol. The order of the subsections indicates the order in which the experiments were executed, in which the best model in a set of experiments was selected as the baseline for the following set of experiments. The last subsection shows the results for a model that considers all the best configurations presented in the preceding subsections.

4.3.1 Learning rate and regularization

Tab. 3 shows the results obtained for the two learning rate configurations: the backbone's learning rate is lower than the model's learning rate ($lr_{backbone} < lr_{model}$), and the same learning rate for the backbone and the model ($lr_{backbone} = lr_{model}$). Moreover, two configurations using regularization techniques are presented: data augmentation and with data augmentation and temporal dropout. The mIoU in the training, validation, and test sets is presented to identify whether overfitting occurs.

Tab. 3: Results obtained for learning rate tuning and regularization techniques (data augmentation and temporal dropout) in terms of mIoU for train, validation, and test sets.

Experiment	mIoU (%) ↑		
	Train	Val	Test
$lr_{backbone} < lr_{model}$	94.2	79.3	79.4
$lr_{backbone} = lr_{model}$	**96.9**	79.9	80.2
w/ data augmentation	89.6	81.2	81.3
w/ data augmentation + temporal dropout	87.1	**81.4**	**81.6**

The default configuration ($lr_{backbone} < lr_{model}$) exhibits overfitting because the training mIoU is very high, but there is a large mIoU drop in the validation set. Using the same learning rate for the backbone and the model achieved higher mIoUs in all sets; however, there was still overfitting owing to the high difference between the training and validation mIoUs. With the inclusion of data augmentation techniques, we can see a mIoU drop in the training set but a mIoU gain in the validation set. Moreover, the gap between the training and validation mIoUs was reduced from 17% to 8.4%, which indicates that overfitting is reduced, improving the model's generalization and performance on new or unseen data. Finally, including temporal

dropout, we see a similar effect to that with data augmentation, where overfitting is diminished even more. The test mIoU followed the same pattern as the validation mIoU, with the model having the same learning rate for the backbone, and the model with data augmentation and temporal dropout the best one with an mIoU of 81.6%. These results show that the model requires a higher learning rate for the spatial encoder than that used in the default configuration by DOFA and strong regularization to reduce overfitting.

4.3.2 Spatial decoders

In the following subsections, we include the number of FLOPs required per component model and the total number of parameters to analyse the model size and its computational complexity and determine the efficiency of each variant.

Tab. 4 shows the results obtained for different spatial decoders in terms of the FLOPs, number of parameters, and mIoU in the test set. DOFA's spatial decoder configuration, with a convolutional neck and UperNet, is the most computationally expensive, with 750 GFLOPs, 60.8 million parameters per decoder for each modality, and approximately 300 million parameters for the whole model. In addition, this configuration is the only one in which the spatial decoder FLOPs are higher than the spatial encoder FLOPs ($FLOPs_{Total} - FLOPs_{Decoder}$). On the other hand, Prithvi's spatial decoder is the least computationally expensive, with almost half of the DOFA's total FLOPs, where the spatial decoder requires only 2.35% of the total FLOPs and achieves the highest mIoU in the test set. The SegFormer decoder obtained the worst mIoU, but it was still less computationally expensive than DOFA and slightly more computationally expensive than Prithvi. Based on these results, we selected Prithvi as the spatial decoder for the following experiments because of its low computational complexity and good performance in the test set, which is a simple and lightweight decoder.

Tab. 4: Results obtained for different spatial decoders in terms of FLOPs, number of parameters, and mIoU in the test set.

Spatial decoder	Head GFLOPs	Decoder GFLOPs	Total GFLOPs	Decoder Params (M)	Total Params (M)	Test mIoU (%)
DOFA (Neck + UperNet)	16.3	401.3	750.3	60.8	295.9	81.6
Prithvi	16.3	**8.4**	**357.3**	**4.7**	**127.7**	**82.0**
SegFormer	16.3	28.3	377.3	**4.7**	127.8	80.8

4.3.3 Data fusion techniques

Tab. 5 shows the results obtained for different data fusion techniques: early, mid, and late fusion, in terms of FLOPs, number of parameters, and mIoU in the test set. In terms of mIoU, all fusion techniques achieved similar results, with a difference of 0.3% between the best and the worst technique, which is not enough to claim that one technique is the best, as these results can change owing to the randomness in each experiment. In terms of total FLOPs and number of parameters, early fusion is less computationally expensive and requires fewer parameters than other fusion techniques. The main reason for this is that in early fusion, many S1 samples are dropped to match the smallest sequence size among all modalities, which is equal to 4 because of the optical sequences. Consequently, a shorter sequence of images is concatenated for early fusion compared with mid or late fusion, requiring fewer FLOPs. Moreover, as many S1 images were dropped in early fusion, there was no drop in mIoU compared with the other techniques using all S1 images, suggesting that long sequences of S1 were not needed or that many S1 images were not relevant. This is supported by the results obtained using late fusion with concatenation and a high temporal dropout for S1 images ($p_{S1} = 0.9$), where the mIoU is similar to that using early fusion and many S1 images were dropped in both techniques. We

selected late fusion with concatenation as the fusion technique because it allows the model to learn features for fusion, which is not possible using early or mid-fusion. Late fusion with concatenation and temporal dropout was not preferred because of the high temporal dropout, which might interfere with temporal feature extraction during training. However, this technique appears promising, and further research will focus on tuning the temporal dropout rate to find a suitable value that reduces the number of S1 images and allows a robust extraction of temporal features.

Tab. 5: Results obtained for different data fusion techniques in terms of FLOPs, number of parameters and mIoU in the test set.

Fusion technique	Spatial Encoder GFLOPs	Decoder GFLOPs	Head GFLOPs	Total GFLOPs	Total Params (M)	Test mIoU (%)
Early	153.1	6.8	16.3	**176.2**	**118.3**	**82.2**
Mid	332.7	6.8	16.3	355.8	**118.3**	81.9
Late: Sum	332.7	8.3	16.3	357.3	127.7	82.0
Late: Concat	332.7	8.3	48.9	390.0	131.3	82.1
Late: Concat ($p_{S1} = 0.9$)	332.7	8.3	48.9	390.0	131.3	**82.2**

4.3.4 Modalities

Tab. 6 shows the results obtained for all possible combinations of modalities: S1, S2, Planet, S1+Planet, S2+Planet, S1+S2, and All (S1+S2+Planet), in terms of FLOPs, total number of parameters, and mIoU in the test set. In terms of the number of parameters, as expected, the most complex model comprises all modalities: S1+S2+Planet, followed by the combinations of two modalities, and finally, by each modality separately.

Tab. 6: Results obtained for different combinations of modalities in terms of FLOPs and mIoU in the test set. S1: Sentinel-1, S2: Sentinel-2, and Planet: PlanetScope.

Modality	Spatial Encoder GFLOPs	Decoder GFLOPs	Head GFLOPs	Total GFLOPs	Total Params (M)	Test mIoU (%)
S1	167.0	22.7	1.8	191.0	**118.3**	61.1
S2	17.3	2.27	1.8	**21.4**	**118.3**	78.6
Planet	148.7	20.4	16.3	185.5	**118.3**	81.1
S1 + Planet	315.4	43.1	32.6	391.1	124.8	81.7
S2 + Planet	166.0	22.7	32.6	221.3	124.8	81.7
S1 + S2	184.0	24.9	3.6	212.5	124.8	79.0
All	332.7	45.3	48.9	427.0	131.3	**82.3**

In terms of the total number of FLOPs, S1 required the highest number of FLOPs because of the higher number of images available in the S1 sequences, followed by Planet because of its high spatial resolution. Using two modalities or all modalities required a number of FLOPs approximately equal to the sum of FLOPs using each modality separately.

Regarding the mIoU in the test set values, the best result was achieved using all modalities, which demonstrates the benefits of multi-modal datasets. Taking a closer look at the effect of each modality, we can observe that S1 obtained the worst mIoU (61.1%), requiring the highest number of FLOPs and being the most inefficient modality. This might be due to redundant information in the long S1 sequences over a short period (growing season), where there were no significant changes between consecutive images over time. In contrast, Planet obtained the highest mIoU with a similar number of FLOPs as S1, followed by S2, which outperformed S1,

requiring approximately 10% of S1's number of FLOPs. When combining the two modalities, the highest improvements were obtained by introducing optical images to the S1 sequences, with mIoU gains of 17.9% (with S2) and 20.6% (with Planet). Using two optical modalities (S2 + Planet) resulted in smaller mIoU gains than when a single modality was used: 3.1% (S2) and 0.6% (Planet). These results demonstrate the benefits of using optical images for LULC classification, which introduce more discriminative features (spectral values) than radar features (polarizations), and the benefits of using both optical and radar images together to further enhance the model, achieving the highest mIoU in the test set.

4.3.5 Temporal encoders

Tab. 7 shows the results obtained for different temporal configurations: single temporal per modality (ST) and with all modalities (ST All), single temporal trained on all modalities and tested on each modality separately (ST All-Modality), multi-temporal (MT), using L-TAE per modality and L-TAE with all modalities (L-TAE All), and using L-TAE only for S1 and temporal max-pooling for S2 and Planet (L-TAE/Max-Pool All). The results are presented in terms of FLOPs, total number of parameters, and mIoU in the test set.

Tab. 7: Results obtained for different temporal configurations in terms of FLOPs and mIoU in the test set. S1: Sentinel-1, S2: Sentinel-2, Planet: PlanetScope. ST: single temporal, MT: multi temporal, ST ALL-Modality: a model trained on all modalities and tested on the specified modality. For L-TAE/Max-Pool All, temporal max-pooling was used for S2 and Planet, and L-TAE for S1.

Temporal configuration	Spatial Encoder GFLOPs	Head GFLOPs	Decoder GFLOPs	Temporal Encoder GFLOPs	Total GFLOPs	Total Params (M)	Test mIoU (%)
ST S1	166.7	54.4	22.7	-	243.8	**118.3**	51.6
ST S2	17.3	5.4	2.3	-	25.0	**118.3**	76.5
ST Planet	148.7	48.9	20.4	-	218.1	**118.3**	80.6
ST All	332.7	587.9	45.3	-	965.9	**118.3**	76.8
ST All-S1	166.7	54.4	22.7	-	243.8	**118.3**	49.3
ST All-S2	17.3	5.4	2.3	-	25.0	**118.3**	70.7
ST All-Planet	148.7	48.9	20.4	-	218.1	**118.3**	78.8
MT S1	167.0	1.8	22.7	-	191.0	**118.3**	61.1
MT S2	17.3	1.8	2.3	-	**21.4**	**118.3**	78.6
MT Planet	149.0	16.3	20.4	-	185.0	**118.3**	81.1
MT All	332.7	48.9	45.3	-	427.0	131.3	**82.3**
L-TAE S1	167.0	1.8	22.7	21.3	212.0	119.0	64.5
L-TAE S2	17.3	1.8	2.3	2.1	23.5	119.0	77.9
L-TAE Planet	149.0	16.3	20.4	19.1	205.0	119.0	81.1
L-TAE All	332.7	48.9	45.3	42.5	469.0	133.4	**82.3**
L-TAE/Max-Pool All	332.7	48.9	45.3	21.3	448.0	132.0	**82.4**

In general terms, MT All, L-TAE All, and L-TAE/Max-Pool All achieved the highest mIoU with a similar number of total FLOPs, with MT All having the lowest computational cost as it uses max-pooling instead of a temporal encoder. It is not possible to select the best temporal configuration based only on the mIoU in the test set, as the highest values are too close to each other and may vary owing to randomness during training.

Regarding the ST models, ST Planet was the best, followed by ST All and ST S2, and ST S1 was the worst. These results confirm that optical images (Planet and S2) are better for this task than radar images (S1), similar to the results in Tab. 6. Then, three methods of using all images in the SITS are explored: using all images from all modalities to train a single model (ST All), max-pooling in the temporal dimension (MT), and learning temporal patterns using a temporal encoder (L-TAE). MT and L-TAE improved the mIoU for each modality separately and using all modalities together, while ST All produced drops in terms of mIoU. This demonstrates the importance of using methods to exploit temporal features from SITS, such as max-pooling (MT) or temporal encoders (L-TAE), instead of using all images without a defined temporal sequence (ST All).
Comparing the improvements between ST per modality and using MT or L-TAE per modality, we observed that L-TAE produced the following mIoU improvements: 12.9% for S1, 1.4% for S2, and 0.5% for Planet. L-TAE leveraged S1 in a better way as S1 was the modality with the longest sequences, and according to GARNOT & LANDRIEU (2021) and GARNOT et al. (2022), L-TAE is usually applied to sequences with 50 or 100 images. This also explains the small improvements for modalities with short sequences, such as Planet, with only 4 images during the growing season. In contrast, MT using temporal max-pooling improved ST S2 mIoU by 2.1%, surpassing L-TAE, which obtained a gain of 1.4%. Thus, we conclude that L-TAE is more suitable for longer sequences to learn temporal patterns, whereas temporal max-pooling is a lightweight option for shorter sequences. For this reason, the proposed temporal configuration L-TAE/Max-Pool All combined both methods by using L-TAE only for long sequences (S1) and max-pooling for short sequences (S2 and Planet); this configuration was selected because of its flexibility and performance.

4.3.6 Weights initialization

Tab. 8 shows the results obtained for different weights initializations for the spatial encoder (ViT-B): training from scratch, using pre-trained weights on ImageNet, and using pre-trained weights from a foundation model (DOFA) in terms of FLOPs, total number of parameters, and mIoU in the test set. As expected, the pre-trained weights achieved better results than those obtained by training from scratch, which demonstrates the benefits of using models pre-trained on large datasets as a starting point for fine-tuning for different downstream tasks. Moreover, using pre-trained weights from a foundation model further improved the mIoU, which confirms that pre-training on a dataset from a similar domain (remote sensing images) is much better than pre-training on a generic domain.

Tab. 8: Results obtained for different weights initialization in terms of FLOPs and mIoU in the test set.

Weights initialization	Spatial Encoder GFLOPs	Head GFLOPs	Decoder GFLOPs	Total GFLOPs	Total Params (M)	Test mIoU (%)
Scratch	332.7	48.9	45.3	448.0	132.0	80.9
ImageNet	332.7	48.9	45.3	448.0	132.0	81.2
DOFA	332.7	48.9	45.3	448.0	132.0	**82.4**

4.3.7 Class-wise analysis

For three different temporal configurations, ST All (single temporal with all modalities), MT All (temporal max-pooling with all modalities), and L-TAE/Max-Pool All (L-TAE for S1, and temporal max-pooling for S2 and Planet), we computed the metrics per class to further analyse the model's performance in recognizing specific LULC classes. Fig. 5 presents the IoU per class and the mIoU in the test set using the three temporal configurations.

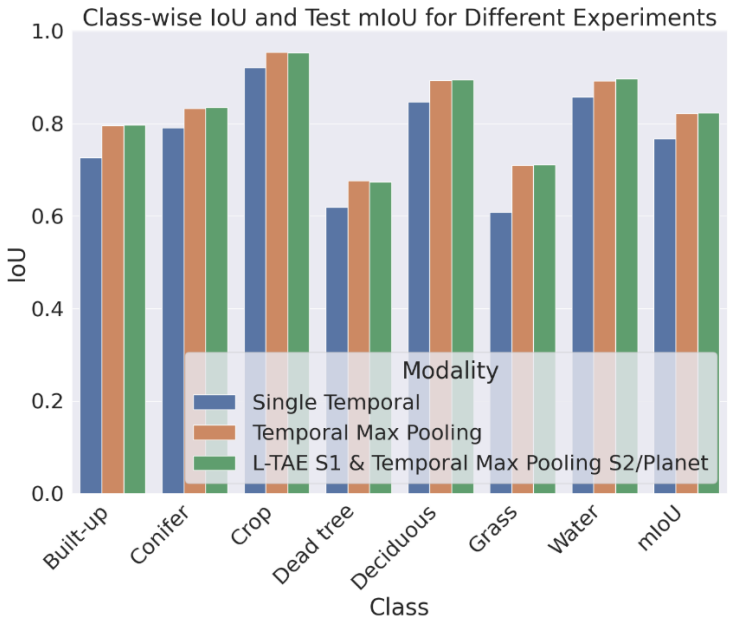

Fig. 5: Class-wise IoU and mIoU obtained using different temporal configurations: ST All (single temporal with all modalities), MT All (temporal max-pooling with all modalities), and L-TAE/Max-Pool All (L-TAE for S1 and temporal max-pooling for S2 and Planet).

In general terms, we observed the same pattern for all classes: L-TAE/Max-Pool All (green bars) and MT All (orange bars) achieved the highest IoU for all classes and mIoU, whereas ST All had the worst performance. Taking a closer look at specific classes, dead trees and grass are the most difficult classes with the lowest IoUs of 0.67 and 0.71, respectively, using the best models. These classes are difficult to classify because of their definitions or similarities to other classes. For instance, based on the visual assessment performed for manual annotation using only very high-resolution optical images (PlanetScope), it is difficult to determine whether a tree is dead, dry (without leaves, even in the growing season), or looks like bare soil. A field trip is needed to further improve the manual annotations and discard potential false positives or look-alikes. Moreover, grasslands are very similar to croplands in their early stages, making it necessary to analyse longer sequences to differentiate them.

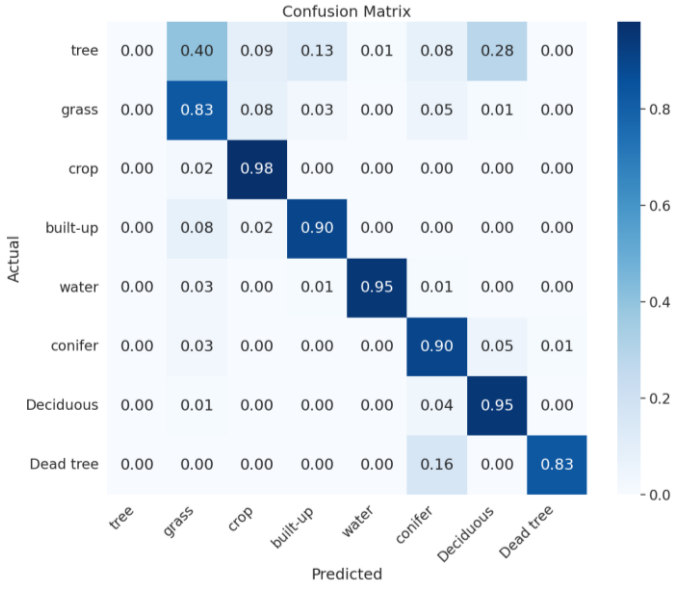

Fig. 6: Confusion matrix obtained using the temporal configuration: L-TAE/Max-Pool All (L-TAE for S1 and temporal max-pooling for S2 and Planet). The values are normalized per row.

We also computed the confusion matrix for the best temporal configuration to identify potential confusion between classes. Fig. 6 shows the confusion matrix obtained using L-TAE/Max-Pool All. Some pixels of the grass class were misclassified as belonging to the crop, built-up, and coniferous tree classes. This was expected for the aforementioned reasons, highlighting the need for longer sequences over time to improve class discrimination. Similarly, many dead tree pixels were misclassified as coniferous trees because it is difficult to determine whether a tree is dead, emphasizing our hypothesis that a visual assessment based only on very high-resolution optical images is not sufficient for the annotation of dead trees. Additional information, such as in situ measurements from field trips, higher spatial resolution images, and even oblique images, can be employed to improve the manual annotations. Finally, the confusion between the grass and built-up classes can be explained by the small backbone patch embedding (16×16 for ViT) and fine-granular labels for the built-up class, requiring an even higher spatial resolution than PlanetScope (3m) to enhance the prediction of this class.

4.4 Qualitative Analysis

In this subsection, we perform a visual assessment of the predictions using the best temporal configuration, L-TAE/Max-Pool All, to analyse the confusion between classes and wrong-annotated classes in our dataset. Because one of the sources for our labels was the ESA LULC Maps, there was uncertainty in the labels because these maps were generated using a machine learning algorithm with an overall accuracy of 74.4% (ESA 2020) and 76.7% (ESA 2021) for 2020 and 2021, respectively. Moreover, manual annotation of dead trees is prone to human error and difficulties in determining whether a tree is dead. Finally, the assumption that the tree species did not change from 2018 to 2020/2021 for the coniferous and deciduous classes also introduces incorrect labels.

Fig. 7 shows some patches where the model predictions were not accurate compared with the reference labels. From left to right: PlanetScope patch from 2020, Sentinel-2 patch from 2020, corresponding ground truth, and model prediction.

Fig. 7a and 7b present the cases for the fine-granular classes: built-up (Fig. 7a) and grass (Fig. 7b). As the ESA labels were created using a machine learning algorithm for pixel-wise classification, the labels were fine-granular (pixel level). This can be seen in Fig. 7a, where built-up areas surrounded by grass are finely delineated in the ground truth, whereas the prediction of the model contains spatially smoothed boundaries between classes, causing confusion between grass and built-up classes. In contrast, Fig. 7b illustrates the case of thin roads and field paths, which were classified as grass in the ground truth and were misclassified by the model as croplands. In this case, even the ground truth was wrong, as the roads and field paths should be assigned to the built-up class, leading to erroneous model training and confusion between those classes.

Fig. 7c highlights a case with incorrect annotations for the dead tree class. The red rectangles indicate areas classified as coniferous trees in the ground truth; however, based on the Planet and S2 images, we can see that those areas should be classified as dead trees. In this case, the model produced a better prediction than the ground truth, despite being trained on inaccurate labels. Note that to discard false positives or reduce confusion between classes, additional information from field trips, higher-resolution images, or longer image sequences is required, as the dead tree class is difficult to distinguish using a single image.

A deforested area is shown in Fig. 7d (red rectangle), which was misclassified as coniferous trees in the ground truth and classified as grass by the model. The model's prediction was more accurate than the ground truth, as this area contained green pixels, which could be related to tree remainders after deforestation. However, the model still failed to predict the entire area as grass, with many pixels misclassified as coniferous trees owing to incorrect labels.

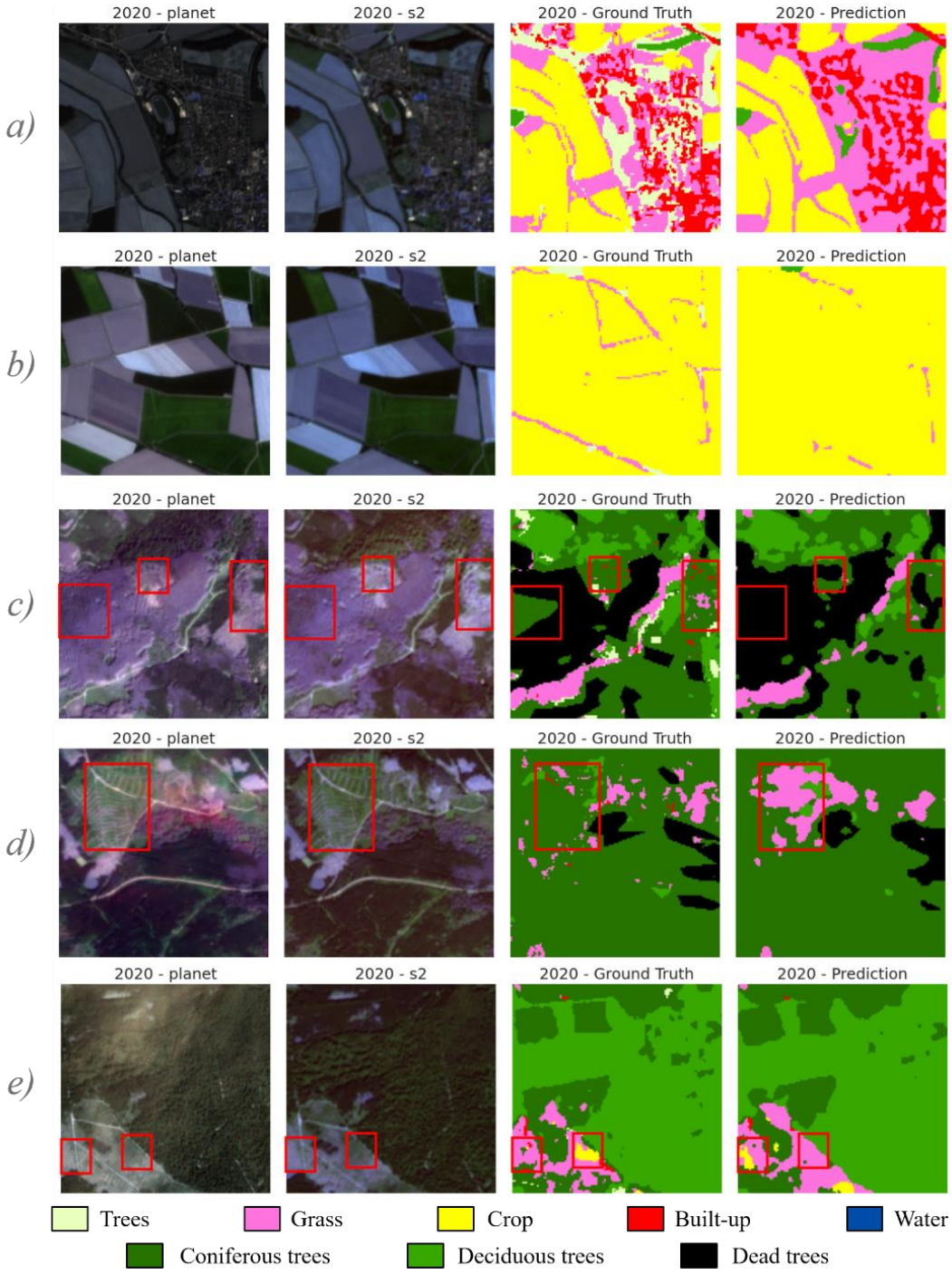

Fig. 7: Comparison between reference labels (ground truth) and our model's predictions using L-TAE/Max-Pool All temporal configuration. From left to right: PlanetScope patch from 2020, Sentinel-2 patch from 2020, corresponding ground truth, and model prediction. From top to bottom: a) and b) predictions for fine-granular classes, c) wrong dead trees annotations, and d) and e) cases with confusion between classes.

Fig. 7e exhibits a case of confusion between the grass and crop classes. In the ground truth, the areas highlighted by red rectangles are classified as grass (left) and crop (right); however, we can see in the Planet and S2 images that both areas look very similar to each other. The model predicted both areas as belonging to the grass class, with a few pixels classified as crops owing to the confusion between those classes. This could be the case when the crop area was in an early stage (germination or seeding) and looked like grass, where we would need a longer image sequence to observe the differences over time.

4.5 Discussion

The regularization methods employed, data augmentation, and temporal dropout successfully reduced model overfitting; however, as there is still a gap between training and validation mIoUs, more regularization methods can be applied. For instance, photometric distortion of spectral images or a more intensive crop & resize transformation with random ratios.

Prithvi's spatial decoder was demonstrated to be a lightweight decoder with better performance than convolutional Neck + UPerNet and SegFormer. Thus, we can conclude that using a spatial decoder with more parameters (Neck + UPerNet) does not necessarily result in higher performance.

Among the data fusion techniques explored in this work, the late fusion technique was preferred as it allows the model to learn features for fusion, which is not possible using early or mid fusion. The addition of temporal dropout only for S1 images to late fusion showed high potential; however, the temporal dropout rate should be carefully tuned to reduce the number of S1 images while maintaining sufficient S1 images to learn robust temporal features. This can be further enhanced by adding group encoding to distinguish features by modalities to temporal encoding based on each image timestamp (CONG et al. 2022), or by using attention to fuse modalities as performed by SkySense and OmniSAT.

Regarding the modalities employed, we noticed that high-resolution optical images (PlanetScope) obtained better results than using only radar images (S1) for LULC classification, with spectral features being more discriminative than radar features (polarizations). However, the computational complexity increased for higher spatial resolution or longer sequences, as expected. In our case, the number of FLOPs for PlanetScope and S1 was similar because of the short sequences with very high-resolution images and the long sequences with a coarser resolution. Furthermore, using all modalities introduced the benefits of both optical and radar features, thereby enhancing the results.

Exploration of temporal configurations allowed us to identify a suitable temporal encoder based on the available sequence length. L-TAE managed to further improve the results and learn temporal patterns in longer sequences, whereas temporal max-pooling is a lightweight option for shorter sequences. The combination of both methods is highly recommended because of their flexibility and performance, especially for cases with different sequence lengths per modality, where L-TAE should be used for longer sequences and temporal max-pooling for shorter sequences.

Using a foundation model such as DOFA (ViT) as a spatial encoder improved the results compared with pre-trained weights on ImageNet or trained from scratch. Other foundation models can be further explored, such as msGFM (HAN et al. 2024), which uses a Swin-Transformer (LIU et al. 2021) with a Mixture of Experts (MoE) (SHAZEER et al. 2017). However, the codes and weights of this foundation model are not yet publicly available.

The best configuration of the proposed framework achieved high IoUs for all classes except for grass and dead trees. These classes were the hardest to classify because of their complex definitions and similarities with the other classes. Using only very high-resolution optical images (PlanetScope) to determine if a tree is dead based on visual assessment is very difficult, as it might look like a dry tree (without leaves) or bare soil. In contrast, grasslands are very similar to croplands in their early stages (seeding or germination). For this reason, a field trip is necessary to discard potential false positives or look-alikes, as well as longer sequences, to distinguish temporal patterns considering other seasons over the year.

After a qualitative analysis of the results, we found that fine granular classes (roads or field paths) were difficult to classify using the proposed model. This is related to the patchifying process performed by the backbone (ViT), in which fine-grained spatial information is discarded, producing a smoother boundary between classes. Moreover, there were many pixels misclassified by the model due to confusion between dead trees, coniferous trees, grass, and

crops. The main reasons for this confusion are incorrect, manually annotated dead trees and uncertainties in labels for other classes. As dead trees were manually annotated, these labels were prone to human error, the annotator's expertise, and difficulty in distinguishing dead trees and look-alikes. For the coniferous and deciduous tree types, the assumption that the tree types did not change from 2018 to 2020/2021 introduced noise to those labels. In addition, the labels derived from the ESA LULC maps contained uncertainties because these labels were generated using a machine learning algorithm with overall accuracies of 74.4% (ESA 2020) and 76.7% (ESA 2021). Further research will focus on the refinement of these labels based on manual inspection or using semi-supervised annotation techniques, such as active teachers (MI et al. 2022).
Our proposed framework uses a shared spatial encoder across all modalities, which may not be the best option for larger discrepancies in spatial resolution. For this reason, recent works, such as SkySense, train separate encoders for very high resolution (0.2m) and multi-spectral Sentinel-2 (10m). This could be introduced into our framework at the cost of increasing the computational complexity of the model; however, the code and weights are not yet publicly available.

5 Conclusions

This study proposes a framework to integrate satellite image time series (SITS) into mono-temporal multi-modal Geospatial Foundation Models (GFMs). Many ablation studies have been performed to assess each component of the proposed framework, such as hyperparameters, regularization techniques, spatial decoders, temporal encoders, data fusion techniques, and the effects of different modalities. The assessment was performed by measuring the computational complexity (in terms of FLOPs and number of parameters) and prediction performance (in terms of mean Intersection over Union (mIoU)). Three modalities were used: Sentinel-1 (radar, 10m), Sentinel-2 (optical, 10m), and PlanetScope (optical, 3m), and DOFA (XIONG et al. 2024), a mono-temporal multi-modal GFM, as a shared spatial encoder across all modalities. As a case study, we used the proposed framework for land use and land cover (LULC) classification in the Harz Mountains region of Germany. We obtained the best results, with a mIoU of 82.4% requiring 448 GFLOPs and 132M parameters, using the same learning rate for the backbone and the entire model, data augmentation and temporal dropout, Prithvi's spatial decoder, late fusion with concatenation, all modalities, and L-TAE (GARNOT & LANDRIEU 2020) for Sentinel-1 and temporal max-pooling for PlanetScope and Sentinel-2.
The use of regularization techniques, such as data augmentation (crop & resize and flips) and temporal dropout (with dropout rates of 0.4 for optical and 0.2 for radar), diminished the high overfitting in the model. Using Prithvi's spatial decoder instead of DOFA's (convolutional neck + UperNet (XIAO et al. 2018)), the computational complexity was reduced, requiring almost half of the DOFA's total FLOPs with a slight improvement in mIoU (0.4%).
Among the fusion techniques, early, mid, and late fusion all achieved similar results in terms of the mIoU. Concerning computational complexity, early fusion required almost half of the FLOPs compared with mid and late fusion. In contrast, late fusion allows the model to learn modality-specific features, especially for long sequences, at a higher computational cost.
Regarding the influence of each modality, we observed that high-resolution optical images from PlanetScope achieved the highest mIoU (81.1%) with an improvement of 1.2% after

including all modalities. On the other hand, Sentinel-1 obtained the worst results with a mIoU of 61.1%, in addition to being the modality with longer image sequences.
The use of temporal encoders is highly recommended for long sequences to learn temporal patterns, whereas temporal max-pooling is a lightweight alternative for short sequences. We merged both approaches to exploit sequences of different lengths. In this way, L-TAE was used for Sentinel-1 sequences, whereas temporal max-pooling was used for PlanetScope and Sentinel-2 sequences.
Based on a visual assessment of the model predictions, we identified two main issues: confusion between classes and inconsistent annotations. Despite these issues that might hinder the learning of the model, there were some cases in which the model managed to produce a more accurate prediction than the corresponding ground truth.
The extensions of this work include the use of different GFMs as spatial encoders, such as msGFM (HAN et al. 2024), spatial encoders per modality, similar to SkySense (GUO et al 2024), additional multi-temporal and multi-modal datasets, and semi-supervised annotation techniques to improve the annotations.

Acknowledgments

The work described in this article was carried out as part of the EXDIMUM project, funded by the Federal Ministry of Education and Research (BMBF), as part of the "WaX - Water Extreme Events" funding programme. The WaX funding measure runs under the umbrella of the federal programme "Wasser: N – Research and Innovation for Sustainability", which was initiated by the BMBF. Wasser:N is part of the Research for Sustainability (FONA) strategy.

6 References

ASTRUC, G., GONTHIER, N., MALLET, C. & LANDRIEU, L., 2024: OmniSat: Self-supervised Modality Fusion for Earth Observation. Proceedings of the European Conference on Computer Vision (ECCV), 409-427, https://doi.org/10.1007/978-3-031-73390-1_24.

BERG, P., UZUN, B., PHAM, M. -T. & COURTY, N., 2024: Multimodal Supervised Contrastive Learning in Remote Sensing Downstream Tasks. IEEE Geoscience and Remote Sensing Letters, **21**, 1-5, https://doi.org/10.1109/LGRS.2024.3385995.

BLICKENSDÖRFER, L., OEHMICHEN, K., PFLUGMACHER, D., KLEINSCHMIT, B., & HOSTERT, P., 2024: National tree species mapping using Sentinel-1/2 time series and German National Forest Inventory data. Remote Sensing of Environment, **304**, 114069, https://doi.org/10.1016/j.rse.2024.114069.

CONG, Y., KHANNA, S., MENG, C., LIU, P., ROZI, E., HE, Y., BURKE, M., LOBELL, D. & ERMON, S., 2022: SatMAE: Pre-training Transformers for Temporal and Multi-Spectral Satellite Imagery. Advances in Neural Information Processing Systems, **35**, 197-211.

DENG, J., DONG, W., SOCHER, R., LI, L.J., LI, K. & FEI-FEI, L., 2009: ImageNet: A large-scale hierarchical image database. Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), 248-255, https://doi.org/10.1109/CVPR.2009.5206848.

DOSOVITSKIY, A., BEYER, L., KOLESNIKOV, A., WEISSENBORN, D., ZHAI, X., UNTERTHINER, T., DEHGHANI, M., MINDERER, M., HEIGOLD, G., GELLY, S. & USZKOREIT, J., 2020: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, https://doi.org/10.48550/arXiv.2010.11929.

EEA, 2022: European Environment Agency – EEA, Economic losses and fatalities from weather- and climate-related events in Europe, https://data.europa.eu/doi/10.2800/530599, letzter Zugriff 10.04.2025.

ESA, 2020: ESA WorldCover 2020, https://worldcover2020.esa.int/, letzter Zugriff 08.04.2025.

ESA, 2021: ESA WorldCover 2021, https://worldcover2021.esa.int/, letzter Zugriff 08.04.2025.

GARIOUD, A., GONTHIER, N., LANDRIEU, L., DE WIT, A., VALETTE, M., POUPÉE, M., & GIORDANO, S., 2023: FLAIR: a Country-Scale Land Cover Semantic Segmentation Dataset From Multi-Source Optical Imagery. Advances in Neural Information Processing Systems, **36**, 16456-16482.

GARNOT, V. S. F. & LANDRIEU, L., 2021: Panoptic Segmentation of Satellite Image Time Series with Convolutional Temporal Attention Networks. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 4872-4881, https://doi.org/10.1109/ICCV48922.2021.00483.

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, https://doi.org/10.1016/j.isprsjprs.2022.03.012.

GARNOT, V.S.F. & LANDRIEU, L., 2020: Lightweight Temporal Self-attention for Classifying Satellite Images. Advanced Analytics and Learning on Temporal Data, LEMAIRE, V., MALINOWSKI, S., BAGNALL, A., GUYET, T., TAVENARD & R., IFRIM, G. (eds), **12588**, 171-181, https://doi.org/10.1007/978-3-030-65742-0_12.

GUO, X., LAO, J., DANG, B., ZHANG, Y., YU, L., RU, L., ZHONG, L., HUANG, Z., WU, K., HU, D. & HE, H., 2024: SkySense: A Multi-Modal Remote Sensing Foundation Model Towards Universal Interpretation for Earth Observation Imagery. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 27672-27683, https://doi.org/10.1109/CVPR52733.2024.02613.

HAN, B., ZHANG, S., SHI, X., & REICHSTEIN, M., 2024: Bridging Remote Sensors with Multisensor Geospatial Foundation Models. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 27852-27862, https://doi.org/10.1109/CVPR52733.2024.02631.

JAKUBIK, J., ROY, S., PHILLIPS, C.E., FRACCARO, P., GODWIN, D., ZADROZNY, B., SZWARCMAN, D., GOMES, C., NYIRJESY, G., EDWARDS, B. & KIMURA, D., 2023: Foundation models for generalist geospatial artificial intelligence, arXiv preprint arXiv:2310.18660, https://doi.org/10.48550/arXiv.2310.18660.

LI, Z., HOU, B., MA, S., WU, Z., GUO, X., REN, B. & JIAO, L., 2024: Masked Angle-Aware Autoencoder for Remote Sensing Images. Proceedings of the European Conference on Computer Vision (ECCV), 260-278, https://doi.org/10.1007/978-3-031-73242-3_15.

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. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 10012-10022, https://doi.org/10.1109/ICCV48922.2021.00986.

LOSHCHILOV, I., & HUTTER, F., 2017: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101, https://doi.org/10.48550/arXiv.1711.05101.

MI, P., LIN, J., ZHOU, Y., SHEN, Y., LUO, G., SUN, X., CAO, L., FU, R., XU, Q. & JI, R., 2022: Active Teacher for Semi-Supervised Object Detection. Proceedings of the IEEE/CVF

Conference on Computer Vision and Pattern Recognition (CVPR), 14482-14491, https://doi.org/10.1109/CVPR52688.2022.01408.

OVERBECK, M. & SCHMIDT, M., 2012: Modelling infestation risk of Norway spruce by *Ips typographus* (L.) in the Lower Saxon Harz Mountains (Germany). Forest Ecology and Management, **266**, 115-125, https://doi.org/10.1016/j.foreco.2011.11.011.

PUTZENLECHNER, B., BEVERN, F., KOAL, P., GRIEGER, S., KAPPAS, M., KOUKAL, T., LÖW, M. & FILIPPONI, F., 2024: Accuracy assessment of LAI, PAI and FCOVER from Sentinel-2 and GEDI for monitoring forests and their disturbance in Central Germany. European Journal of Remote Sensing, **57**(1), https://doi.org/10.1080/22797254.2024.2422323.

SHAZEER, N., MIRHOSEINI, A., MAZIARZ, K., DAVIS, A., LE, Q., HINTON, G. & DEAN, J., 2017: Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. arXiv preprint arXiv:1701.06538, https://doi.org/10.48550/arXiv.1701.06538.

TSENG, G., CARTUYVELS, R., ZVONKOV, I., PUROHIT, M., ROLNICK, D., & KERNER, H., 2023: Lightweight, Pre-trained Transformers for Remote Sensing Timeseries. arXiv preprint arXiv:2304.14065, https://doi.org/10.48550/arXiv.2304.14065.

UFZ, 2023: Drought Monitor Germany. Helmholtz Centre for Environmental Research GmbH – UFZ Leipzig, https://www.ufz.de/index.php?en=37937, letzter Zugriff 10.04.2025.

XIAO, T., LIU, Y., ZHOU, B., JIANG, Y., & SUN, J., 2018: Unified perceptual parsing for scene understanding. Proceedings of the European Conference on Computer Vision (ECCV), **11209**, 432-448, https://doi.org/10.1007/978-3-030-01228-1_26.

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.

XIONG, Z., WANG, Y., ZHANG, F., STEWART, A.J., HANNA, J., BORTH, D., PAPOUTSIS, I., SAUX, B.L., CAMPS-VALLS, G. & ZHU, X.X., 2024: Neural plasticity-inspired multimodal foundation model for Earth observation. arXiv preprint arXiv:2403.15356, https://doi.org/10.48550/arXiv.2403.15356.

ZHAO, Y., LIU, J., YANG, J. & WU, Z., 2022: Remote Sensing Image Scene Classification via Self-Supervised Learning and Knowledge Distillation. Remote Sensing, **14**(19), 4813, https://doi.org/10.3390/rs14194813.