ISPRS Journal of Photogrammetry and Remote Sensing 232 (2026) 578–593

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

Mapping land uses following tropical deforestation with location-aware deep learning

Jan Pišl [a,*], Gencer Sumbul [a], Gaston Lenczner [a], Camilo Zamora [b], Martin Herold [b], Jan Dirk Wegner [c], Devis Tuia [a]

[a] École Polytechnique Fédérale de Lausanne (EPFL), Sion, Switzerland
[b] GFZ Helmholtz-Zentrum Potsdam, Germany
[c] University of Zurich, Switzerland

ARTICLE INFO

Keywords:
Remote sensing
Deforestation
Deep learning
Location encoding
Time series
Multi-modal classification

ABSTRACT

The rates of tropical deforestation remain alarmingly high. To enable effective, targeted policy responses, detailed data on its driving forces is needed—each deforestation event needs to be attributed to an agricultural commodity or another land use. Remote sensing allows us to monitor land use conversion following deforestation, providing a proxy of drivers. However, recognizing individual commodities is challenging due to spectral similarities, the limited spatial resolution of free satellite imagery, and limited labeled data. To tackle these challenges, we propose a deep learning, multi-modal approach for the recognition of post-deforestation land uses from a time series of Sentinel-2 images, geographic coordinates, and country-level statistics of deforestation drivers. To integrate the modalities, we design a Transformer-based model with modality-specific encoders. The approach reaches 87% accuracy, an improvement of 10% over the image-only baseline, with little increase in data volume, computations, and model size. It works well in low-data regimes, and can be easily extended to include other modalities. Overall, this work contributes towards detailed, repeatable, and scalable mapping of deforestation landscapes, providing necessary data for the design and implementation of targeted interventions to protect tropical forests.

1. Introduction

Tropical deforestation – the conversion of forests to other land uses – is a major environmental issue with global consequences. Tropical forests are epicenters of biodiversity, hosting more species than any other ecosystem (Brown, 2014). They play an important role in mitigating climate change due to their capability to store carbon and provide ecosystem services such as food, fuel, shelter or medicine to over one billion people (Lewis et al., 2015). Yet, 95% of all deforestation worldwide, which is estimated at 10 million hectares annually (FAO, 2020), takes place in the tropics (Curtis et al., 2018).

In the past decade, major improvements have been made towards detecting and mapping changes in tropical forests with satellite imagery. Datasets such as the Global Forest Change (GFC) (Hansen et al., 2013) or the Tropical Moist Forest (TMF) (Vancutsem et al., 2021) offer high-resolution, annually updated data on forest canopy change going back several decades, and deforestation alert systems provide near-real-time detections of forest canopy disturbances (Hansen et al., 2016; Mullissa et al., 2023; Tang et al., 2023).

Despite their importance in monitoring tropical forests, detection alerts are not sufficient on their own. To design effective policy responses to tackle deforestation, it is crucial to identify the causes and motivations behind it, also known as *drivers* (Curtis et al., 2018; Geist and Lambin, 2002; Kissinger et al., 2012). Two types of deforestation drivers are generally considered: proximate and underlying (Geist and Lambin, 2002). This work is concerned with proximate drivers, which are human activities that directly cause deforestation (agricultural production, mining, etc.). Underlying drivers correspond to broader political, economic, or demographic forces that impact deforestation indirectly, such as population migration or increased demand for certain goods. Drivers can also be considered at different levels of granularity. They are often grouped into categories such as ‘agriculture’, but the knowledge of specific commodities is important since each driver has a different impact on carbon emissions, biodiversity, and local population (Shapiro et al., 2023).

The attribution of deforestation to specific drivers is not trivial. The term *driver* implies causality (Bernhard et al., 2024), which is difficult

* Corresponding author.
E-mail address: jan.pisl@epfl.ch (J. Pišl).

https://doi.org/10.1016/j.isprsjprs.2025.12.007
Received 17 February 2025; Received in revised form 6 December 2025; Accepted 12 December 2025
Available online 7 January 2026

to prove. Instead, *post-deforestation land use* (land use established on the deforested land) is widely used as a proxy of drivers (Hosonuma et al., 2012; Irvin et al., 2020; Masolele et al., 2024). While the two terms are often used interchangeably, post-deforestation land use analysis does not provide direct evidence that the land use drove (was the motivation for) the forest clearance. Nevertheless, in the absence of robust and scalable methods for determining causality in deforestation dynamics, considering subsequent land use as a proxy remains the standard approach to attribute deforestation to drivers (Geist and Lambin, 2002).

The current knowledge of post-deforestation land uses is limited. Globally, it is well known that most tropical deforestation is followed by agricultural production (Kissinger et al., 2012; Curtis et al., 2018; Masolele et al., 2021), and that the seven *forest-risk commodities* – beef, soy, palm oil, rubber, cocoa, coffee and wood fiber – are associated with the majority of tropical deforestation (Goldman et al., 2020; Pendrill et al., 2022). However, spatially explicit and commodity-specific data are still lacking in many parts of the world (Goldman et al., 2020; Pendrill et al., 2022).

Modeling post-deforestation land uses from satellite images has the potential to fill this gap as it brings several key advantages: it is spatially and temporally explicit, allows to produce global maps at a high resolution, and can be automated with machine learning. The recognition of land use classes with distinct visual patterns, such as urban infrastructure, is relatively straightforward. Recognizing individual commodity types that often have similar spectral responses is more difficult, especially considering the spatial resolution (10–30 m) of freely available satellite imagery produced by the Sentinel-2 or Landsat sensors.

To address this problem, time series of satellite images are frequently used in crop mapping, since they can capture phenological patterns (Rußwurm and Korner, 2017; Campos-Taberner et al., 2020). This has also been adopted in the analyses of post-deforestation land use (Pisl et al., 2024b; Masolele et al., 2021), but the heavy cloud cover typical for the tropics limits its potential. Other data sources, or *modalities*, have therefore been explored to complement satellite images for monitoring the dynamics of post-deforestation land use.

In this work, we present a deep learning-based, multi-modal approach to map post-deforestation land uses, including the seven *forest-risk commodities* at the pan-tropical level. We aim to overcome the difficulty of recognizing land uses from satellite images alone by considering two other data modalities that provide the model a notion of location.

First, we use geographic coordinates as an additional modality. Over the course of training, this allows the model to associate certain classes with locations where they often appear. This can help recognize post-deforestation land uses that do not have a distinct visual appearance, and therefore improve the model performance.

Second, we include country-level data from the DeDuCE dataset (Singh and Persson, 2024). DeDuCE is a tabular dataset that estimates the attribution of deforestation to individual agricultural commodities at a national scale. While the spatial scale is coarse, DeDuCE is available for all countries and most agricultural commodities. Therefore, we hypothesize that it can serve as a prior, indicating which classes are more or less likely to appear in a given country. We expect that it complements the other two modalities and further improves the model performance.

Our approach utilizes the fact that many of the post-deforestation land uses – especially those related to agriculture – are strongly clustered in space. For example, the majority of deforestation in Brazil is followed by establishing pasture for cattle (Pendrill et al., 2022), while this is rare in Indonesia and Malaysia, where forests are most often replaced by oil palm plantations (Goldman et al., 2020). Therefore, we hypothesize that providing the model with a notion of location through additional modalities can enable it to learn such clustering and consequently improve its performance. Importantly, both of these modalities are freely available for any location in the tropics.

To fuse the different modalities, we design a Transformer-based model architecture with modality-specific encoders. To train the model, we create our own training dataset which we release publicly in a ready-to-use format.

Our main contributions are (i) a new approach for the classification of post-deforestation land use from free and openly available multi-modal data, fused by a custom deep learning model; including these modalities allows for better, more accurate fine-grained mapping of the post-deforestation land uses than is currently available, with little overhead in terms of data and computations; (ii) a training and evaluation dataset, named PLUTo (“Post-deforestation Land Use in the Tropics”), compiled from 19 individual datasets covering 11 classes and over 60 countries; we publish the dataset to facilitate further work on data-driven mapping of post-deforestation landscapes.

2. Related works

While this work is concerned with mapping land uses to estimate deforestation drivers, we review works on mapping drivers of any forest loss. In addition to deforestation this also includes temporary disturbances, such as wildfire, which are followed by forest regrowth rather than a land use conversion. The distinction is important for environmental and policy reasons, but the methodology is similar — satellite images and auxiliary data are analyzed to identify indicators of disturbance types, such as burn scars after a fire.

2.1. Automatic mapping of drivers

Most studies employ machine learning to automate this task, since manual classification of forest loss events (Tyukavina et al., 2018; De Sy et al., 2019; Laso Bayas et al., 2022; Fritz et al., 2022) is time-consuming and poses limits on the sample size. Earlier works rely on decision tree-based models (Hermosilla et al., 2015; Richards and Friess, 2016; Schroeder et al., 2017; Nguyen et al., 2018; Alonso et al., 2022), trained on features extracted from satellite images – such as spectral indices – and other data sources including population data, road network, elevation data, and fire alerts. Temporal features, such as the duration of the forest disturbance or the evolution of spectral indices in time, have also been used (Hermosilla et al., 2015; Schroeder et al., 2017; Nguyen et al., 2018).

Notably, Curtis et al. (2018) produced the only global forest loss attribution dataset to date, assigning the dominant forest loss driver for each 10 × 10 km cell, using a random forest model. However, the spatial scale is coarse, it recognizes only five broad classes, and the map is valid with respect to 2001–2015 and has not been updated.

2.2. Deep learning-based methods

Recent approaches are increasingly based on deep learning (Irvin et al., 2020; Mitton and Murray-Smith, 2021; Masolele et al., 2021, 2022; Pisl et al., 2024b), where models learn rich features directly from input data. Architectures designed specifically for learning from images, such as convolutional neural networks (CNNs), can learn complex spatial patterns, resulting in superior performance compared to traditional machine learning (Krizhevsky et al., 2012). For the task at hand, where satellite images play a key role, this is a crucial advantage.

Irvin et al. (2020) are the first to apply deep learning for classifying forest loss drivers, recognizing four driver classes across Indonesia. They achieve a 13% improvement in accuracy compared to a random forest baseline. This work is extended by Mitton and Murray-Smith (2021), improving the model performance with rotation-equivariant convolutions, and Kaselimi et al. (2022), who investigate the potential of a Vision Transformer, showing it performs on par with CNNs. While these approaches show the benefits of using deep learning, they are limited to a single country and only four broad classes.

Multiple works investigate the use of time series of satellite images (Masolele et al., 2021, 2022; Pisl et al., 2024b). Masolele et al. (2021) compare four spatio-temporal architectures against spatial- and temporal-only baselines, obtaining the best performance with a combination of convolutional and attention layers. Pisl et al. (2024b) propose an architecture combining convolutional, recurrent, and attention layers. They show their model learns different strategies for each class, with time series being particular important for the recognition of agriculture-related land uses. Both works conclude that using multi-temporal data brings a significant advantage over using single images if a dedicated spatio-temporal model architecture is used. Masolele et al. (2022) focus instead on the input data, analyzing the benefit of time series for three multi-spectral sensors, Planet, Sentinel-2 and Landsat-8. They observe a significant improvement when using Sentinel-2 images, but minimal with data from the other two sensors.

2.3. Multi-modal fusion

The integration of multi-modal data (Irvin et al., 2020; Masolele et al., 2022; Slagter et al., 2023) is another promising research direction. This includes combining image data from different satellite sensors, topographic, climatic or population data.

There are three general categories of how multi-modal data can be integrated. In *early fusion*, the modalities are combined before the input layer. This is a common scenario in remote sensing, when data from multiple satellite sensors is concatenated along the channel dimension, and treated as a multi-band image. *Mid-fusion* refers to the fusion of representations of individual modalities learned through separate model branches, followed by shared layers enabling interactions between the modalities. Finally, in *late fusion*, predictions are made independently with a separate models for each modality, and the output predictions are combined together.

Recently, mid-fusion has been increasingly recognized for its ability to learn rich intermediate representations of individual modalities and then jointly optimize their combination (Mena et al., 2024; Li et al., 2022). In contrast to early fusion, it allows to preserve inductive biases to individual modalities (such as convolutional layers for image data), and better handles shape and size mismatches between the modalities. Crucially, mid-fusion also enables cross-modal interactions in the feature space, which is its main advantage over late fusion.

While mid-fusion can be implemented through different architectures, a key trend is the use of the attention mechanism, such as the one in the Transformer (Mena et al., 2024). This approach has been used across various remote sensing applications, such as land cover mapping (Liu et al., 2022; Roy et al., 2023) or agriculture (Peng et al., 2024). It allows to learn dynamic interactions between modalities, which often results in superior performance in comparison to using fully-connected (FC) layers and other commonly used methods (Peng et al., 2024; Roy et al., 2023). Other formulations of mid-fusion have been used depending on the specifics of the input modalities, such as graph neural networks for vector data (Pastorino et al., 2022). For a detailed overview of multi-modal fusion in remote sensing, we refer the reader to Li et al. (2022).

The results show that multi-modal data is beneficial particularly in scenarios where individual modalities face limitations. For example, incorporating Sentinel-1 data proves to be especially valuable under conditions of extreme cloud cover (Garnot et al., 2022; Slagter et al., 2023). As discussed above, medium resolution optical images such as those from Sentinel-2 might not be sufficient to recognize individual agricultural commodities on their own, so these findings emphasize the importance of exploring relevant data sources to complement them.

Overall, the methods for attribution of deforestation to drivers from satellite images are improving at a fast pace, largely thanks to advances in deep learning. However, there are still no models that can accurately recognize the main commodities and land uses associated with deforestation across the tropics in a fully automatic way. Existing approaches have been limited to single regions (Irvin et al., 2020; Mitton and Murray-Smith, 2021; Masolele et al., 2022) or to smaller sets of broader, more general land use categories (Masolele et al., 2021; Pisl et al., 2024b).

In our work, we combine the two promising directions, multi-modal fusion and time series. Time series have been shown to consistently improve performance over different models, regions, and output classes. In contrast, the multi-modal fusion arguably has not reached its full potential, which we attribute to the selection of input modalities. We train our model on a new dataset, enabling the recognition of the eleven major classes associated with deforestation including the *forest-risk commodities* across tropics.

3. Data

In this work, we compile and publish the Post-deforestation Land Use in the Tropics (PLUTo) dataset. It can be accessed and downloaded at https://zenodo.org/records/17831353.

PLUTo consists of locations deforested between 2000 and 2020 with known post-deforestation land use. To identify deforestation, we use the TMF dataset (Vancutsem et al., 2021) that provides high-resolution, long-term monitoring of all tropical moist forests globally from 1990. The dataset is produced based on Landsat imagery and maps annual changes in forest, such as deforestation, degradation and regrowth. We only use the annual deforestation layer. The method used for the creation of TMF was designed specifically for tropical moist forests, and is therefore more accurate in comparison to similar global datasets. Further, in contrast to the widely used GFC dataset (Hansen et al., 2013; Tropek et al., 2014; Hansen et al., 2014), TMF excludes certain types of changes in forest canopy, such as regular clearings of forest plantations, or agricultural practices such as replanting. These events do not correspond to deforestation (the conversion of forest to other land uses), and therefore should not be included in the analysis of post-deforestation land use.

The PLUTo dataset consists of 14755 samples, and its design is inspired by the crowd-sourcing campaign of Laso Bayas et al. (2022). We consider a set of 11 post-deforestation classes. Examples of one satellite image per class are given in Fig. 1. Each sample corresponds to a 1 km^2 patch that has experienced deforestation in the target time period and consists of:

- a time series of four Sentinel-2 satellite images
- a deforestation mask extracted from TMF
- the location of the center of the images
- a set of 11 numerical values (one per class) describing the deforestation attributed to each class by the DeDuCe dataset per country; a placeholder for missing data is added if the value is not available
- one land use class, used as a response variable

Examples of the samples including all input modalities and the model's predictions are provided in Fig. D.11.

The target classes include land uses corresponding to the seven *forest-risk commodities* - cattle, oil palm, soy, cocoa, coffee, wood fiber and rubber. These commodities have driven the majority of tropical deforestation in the past decades (Pendrill et al., 2022; Goldman et al., 2020). The remaining classes are *shifting agriculture*, *mining*, *buildings/roads*, and *other*.

To create the dataset, we identify all 1 km^2 patches with substantial deforestation (a minimum of 15 TMF pixels, approximately 1.35 ha) in the target period. We overlay these candidate locations with thematic datasets to identify the post-deforestation land use, which is used as a label.

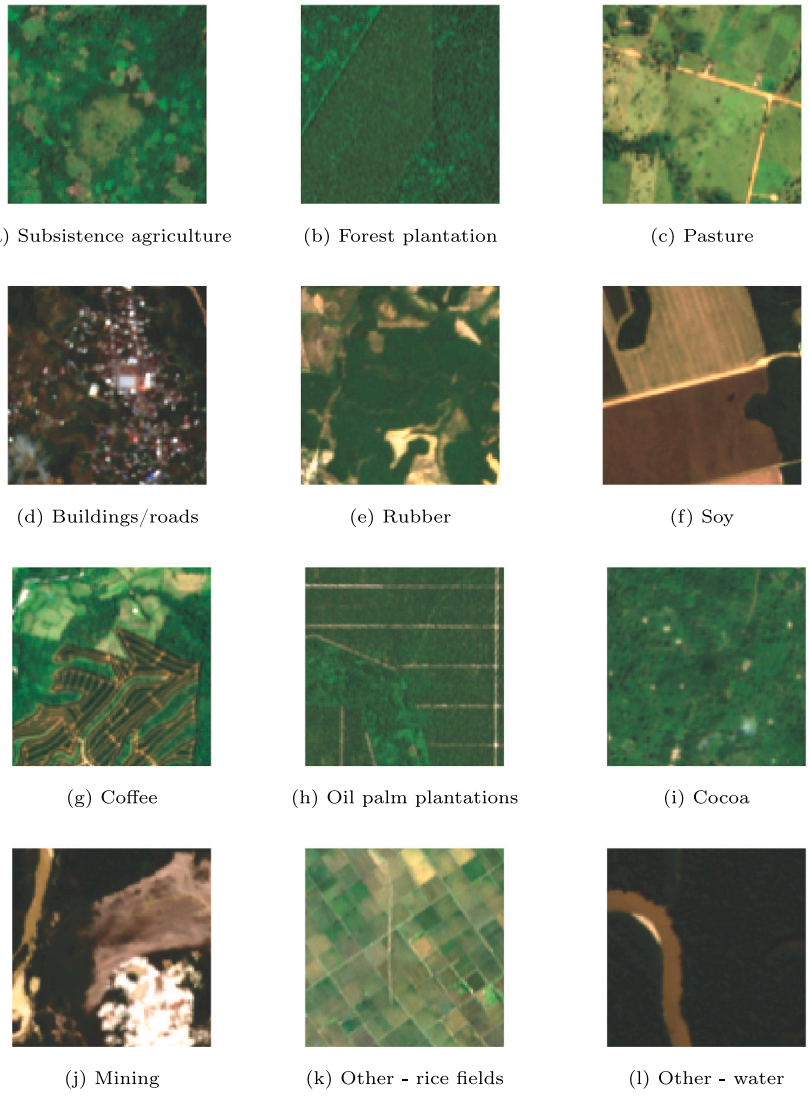


Fig. 1. Examples of classes used in this work.

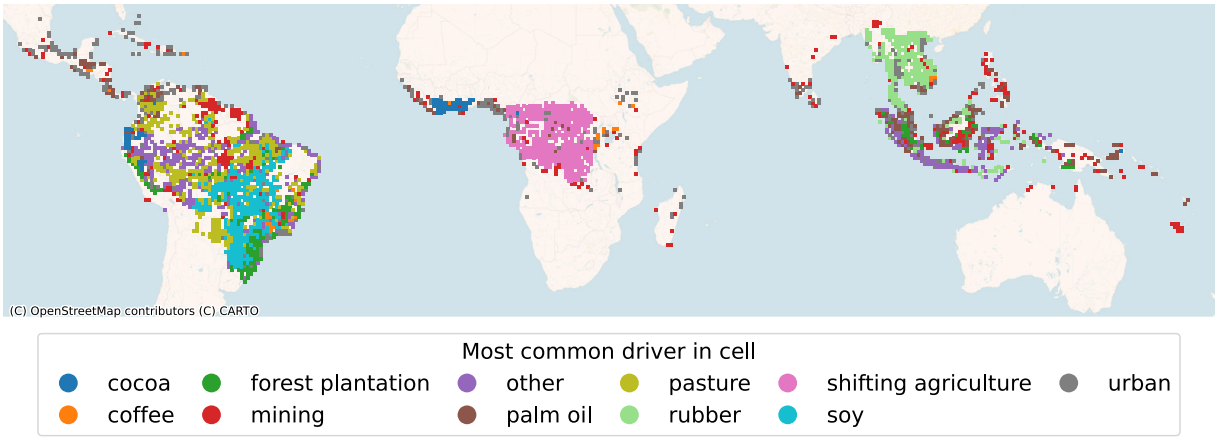
The thematic datasets we used include maps of individual agricultural commodities (Wang et al., 2023; Descals et al., 2020; Kalischek et al., 2022; Maskell et al., 2021; Petersen et al., 2016; Song et al., 2021; Ceccarelli et al., 2021; Becerra et al., 2022), datasets from the MapBiomas project in countries where it is available (MapBiomas Brazil, 2024; MapBiomas Indonesia, 2024; MapBiomas Paraguay, 2024; MapBiomas Bolivia, 2024; MapBiomas Peru, 2024), the Worldcover land cover map for *buildings/roads* (Zanaga et al., 2022), one dataset of smallholder agriculture of the Congo basin Tyukavina et al. (2018), global database of mining polygons (Maus et al., 2022), as well as several datasets covering Colombia, Ecuador and Peru, obtained either through national geoportals or email communications with the representatives of the geoportals. For the class *other*, we sample all other classes from the MapBiomas datasets including different crops (e.g., rice, sugar cane, cotton), and other land covers (water, sand).

We include a candidate location in the training dataset if (i) one or more thematic datasets cover the given location, (ii) there is no conflict between thematic datasets (in case multiple datasets cover the location), and (iii) there is a single class that covers the majority of the deforested area (a threshold of minimum 50% of deforested pixels was used).

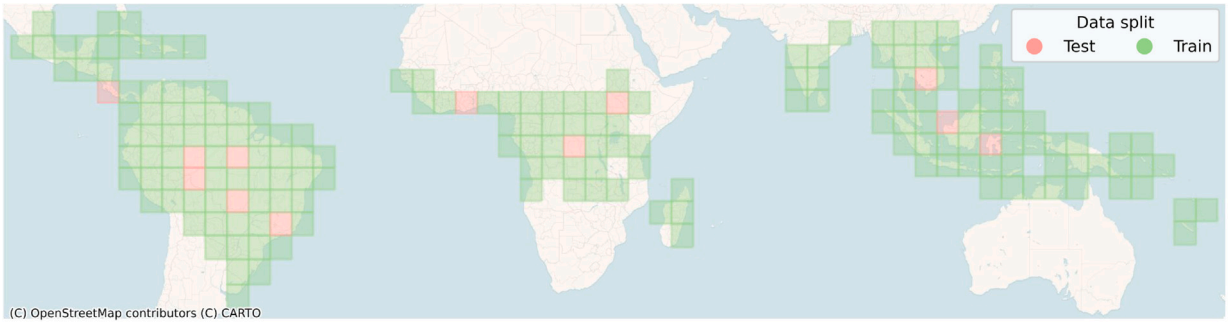
Table 1
Distribution of the samples of the classes of the PLUTo dataset.

Class	Train	Test	Total
Cocoa	938	1113	2051
Coffee	575	198	773
Forest plantation	952	209	1161
Mining	941	532	1473
Oil palm	934	416	1350
Other	933	429	1362
Pasture	945	422	1367
Rubber	945	455	1400
Shifting agriculture	948	281	1229
Soy	944	441	1385
Buildings/roads	945	259	1204
Total	10 000	4755	14 755

We split the dataset into train and test sets as shown in Fig. 2(b). We divide the study area into blocks of 5×5 degrees and select 12 blocks as the test set. We select them in such a way that the country of Cambodia



(a) Distribution of the samples



(b) Geospatial train/test data split

Fig. 2. The Post-deforestation Land Use in the Tropics (PLUTo) dataset.

is entirely within a test block. This allows to test the capabilities of the model to generalize to new, unseen tabular data values in a country unseen in training. All training blocks are further split into 4 groups and used for training and validation using 4-fold cross validation. The class distribution of the dataset is shown in Table 1.

Images. We use time series of Sentinel-2 from the year 2023, for all deforestation that occurred in the target period 2000–2020. We choose 2023 to allow at least three years between the deforestation events (which took place at the latest in 2020) and the acquisition date of images. It may take up to three years before some classes, especially oil palm plantations, have grown enough to be recognizable from satellite images (Goldman et al., 2020). One example of image per class is provided in Fig. 1.

For each quarter of the year, we download the image with the least cloud coverage. We use the L2 A product (surface reflectance) and keep the four bands (RGB+NIR) available at 10 m resolution.

Deforestation mask. We extract the deforestation mask from TMF, covering the same 1 km^2 area as the images. Each pixel of TMF either contains the year of deforestation, or 0 if no deforestation was detected. After preprocessing (Fig. 3), the mask is used as an additional input in the image encoder (described in Section 4.1) of our model.

Location. The location consists of a pair (latitude, longitude), corresponding to the center of the patch.

Country-level tabular data. We use the Deforestation Driver & Carbon Emission (DeDuCE) dataset (Singh and Persson, 2024), which estimates the attribution of deforestation to 184 commodities in 179 countries over the period 2001–2022. To do that, the authors first combine the GFC dataset on forest loss with maps of agricultural commodities, land use and deforestation drivers, to attribute deforestation spatially to broad land use categories, such as plantation or cropland. In the second step, they use tabular data on the expansion of the harvested area of individual crops, such as those provided by FAOSTAT (Food and Agriculture Organization of the United Nations (FAO), 2025), to disaggregate the deforestation attribution to individual agricultural commodities. As a result, their dataset provides estimations on how much deforestation (in hectares) has been driven by each commodity for each country and year.

Despite the coarse spatial granularity of this data, we believe that it can serve as a valuable source of information to the model because of the strong regional patterns. For example, according to DeDuCE, the deforestation attributed to oil palm plantations in Indonesia comprised over 52% of all deforestation in the country. In the Philippines, deforestation is dominated by forest plantations and rice cultivation, and oil palm plantations have driven only 1% of the deforestation there. Considering the visual similarity of forest plantations and oil palm plantations on satellite images (Fig. 1), providing this information to the model might be crucial in order to make the correct prediction.

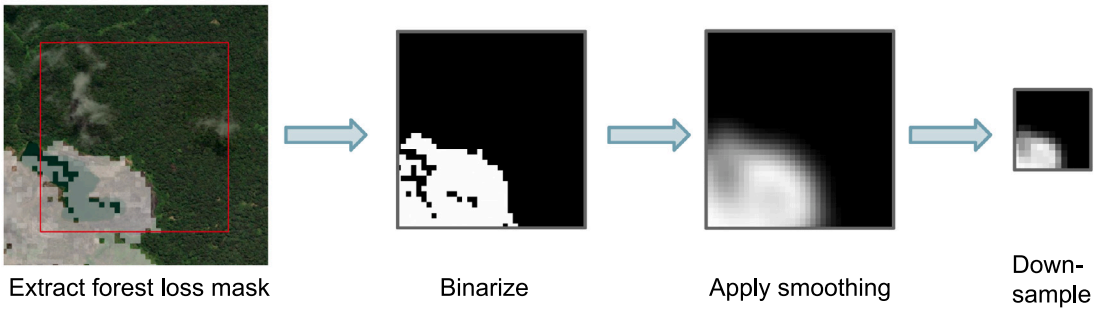


Fig. 3. Preprocessing of the deforestation mask that is used as an input to the model.

To serve as input to the model, the data was normalized by subtracting the mean and dividing by the standard deviation, computed over the entire training dataset. Since the DeDuCE data is available for each year, we used TMF to identify the earliest and latest deforestation year that occurs in a given deforested patch. Then, we averaged the DeDuCE data for each year in that period.

The input to the model therefore consists of 11 numerical values, each representing the relative amount of deforestation attributed to each class for a given country in a given period of time. This corresponds to a prior probability distribution over the target classes-a higher value indicates that given class is relatively more likely to be the deforestation driver in a given location.

For most countries, the data is available for all seven forest-risk commodities we recognize in our work. Where data is missing, we replace it with a learnable scalar indicating missing data. The value of this scalar is optimized as a parameter during training. The missing data token is always used for classes *mining*, *buildings/roads*, *shifting agriculture*, and *other*, because DeDuCe does not provide data for these classes. For patches intersecting multiple countries, we use the tabular data for the country with the largest intersection.

4. Method

This section details our approach for the classification of post-deforestation land uses from multi-modal data. The model architecture is summarized in Fig. 4.

Each input data point $x = \{x_{\text{TS}}, x_{\text{def}}, x_{\text{loc}}, x_{\text{tab}}\}$ consists of a time series x_{TS} of satellite images, a mask x_{def} from TMF indicating on which parts of the satellite images deforestation took place, the location x_{loc} of the center of the images, and a set x_{tab} of 11 numerical values (one per class). Each modality is encoded into feature vectors with a separate encoder, with the exception of x_{def}, which is used as an additional input to the image encoder. We refer to the feature vectors as tokens, and set their size z to 128. All tokens are then processed by a shared Transformer encoder. We choose to use the Transformer since it has established itself as the state-of-the-art architecture for multi-modal fusion across various domains (Radford et al., 2021; Girdhar et al., 2023), including remote sensing (Roy et al., 2023; Aleissaee et al., 2023). The classification token produced by the Transformer is passed to a FC classifier to make a prediction. In the following sections, we describe each part of the network in detail.

4.1. Image encoder

The image encoder $Enc_{\text{img}}(I_i; \theta_{img}) : \mathbb{R}^{w \times h \times c} \rightarrow \mathbb{R}^z$ extracts a token t_i for each satellite image I_i from the input time series x_{TS}. This results in a set of tokens $T_{img} = \{t_i\}_{i=1}^n$, where n denotes the number of the input images. We use a ResNet18 (He et al., 2016) as a backbone, however our approach does not depend on the choice of a specific backbone.

Deforestation mask. To guide the image feature extractor to focus on specific parts of the images, we use the deforestation mask x_{def} for the area covered by the input satellite images. We do this because the deforested area covers only a part of the image. The rest of the image may still contain relevant context, but the prediction should be made mainly with respect to the deforested area.

To use the deforestation mask as input, we first pre-process it as shown in Fig. 3. We binarize it to contain 1 if there was deforestation in given pixel within the target period, and 0 otherwise. We then apply Gaussian blurring and downsample it to the size equal to the dimension of the one of the feature maps of the image encoder. We then multiply pixel-wise the pre-processed mask with the feature map extracted from the image. We choose the 3rd feature map in the ResNet18 image encoder empirically as it yields the best performance on the validation set.

4.2. Location encoder

For every input sample, x_{loc} is provided as a pair of latitude and longitude coordinates. The location encoder $Enc_{\text{loc}}(x_{loc}, \theta_{loc}) : \mathbb{R}^2 \rightarrow \mathbb{R}^z$ transforms these two input values into a z-dimensional token representation, useful for the downstream task. Typically, location encoders consist of a non-parametric positional embedding (PE) and a parametrized neural network (NN) (Rußwurm et al., 2023).

Transforming the raw coordinate values with a PE is desirable to avoid discontinuities on the dateline. Our dataset spans most of the tropics, therefore we expected a continuous representation of the coordinate values to be beneficial. Since each PE method embeds the geographic context in a different way, we evaluated several PEs to find the one that achieves the best performance. The selection of PEs was inspired by Rußwurm et al. (2023). We did not include PEs with tunable hyperparameters to reduce the complexity of the comparison.

As for NNs, we compare a single linear layer as a baseline, FCNet (Mac Aodha et al., 2019) as it is commonly used for location encoding (Rußwurm et al., 2023), and SIREN, Sitzmann et al. (2020), which is particularly suited for representing location given its sinusoidal activation function. Similarly to comparing PEs, we aim to empirically test which NN achieves the best performance.

In total, we compare 9 variants of location encoding, each consisting of a PE and NN applied to the input pair of coordinates. The selection of PEs and NNs was inspired by Rußwurm et al. (2023), and the selected methods are described in Appendix A.

4.3. Tabular encoder

The tabular encoder, $Enc_{\text{tab}}(x_{tab}, \theta_{tab}) : \mathbb{R}^{11} \rightarrow \mathbb{R}^z$ is a single fully connected layer, followed by a ReLU activation function, that maps the 11 input values into a z-dimensional token.

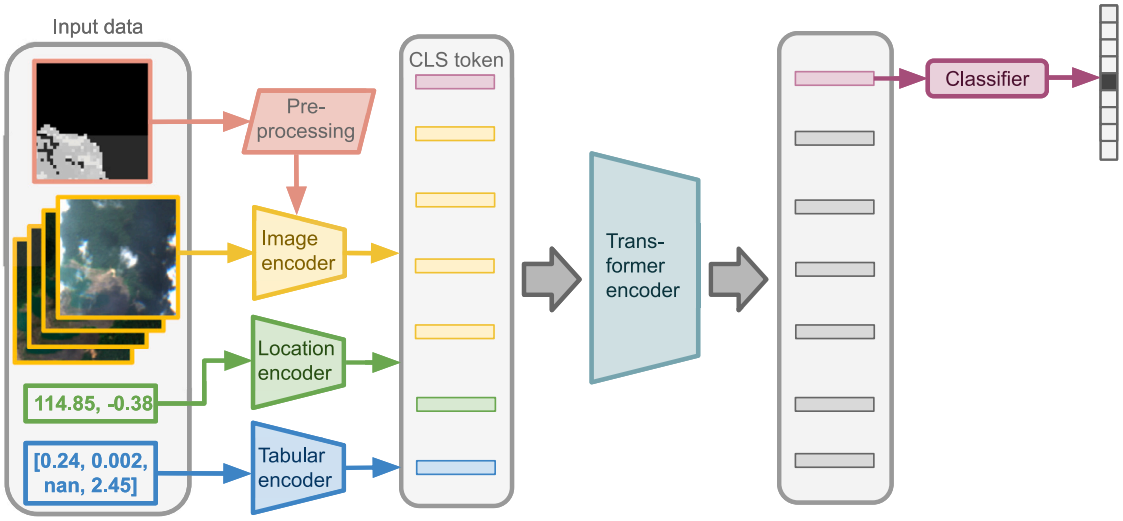


Fig. 4. Transformer-based architecture for multi-modal post-deforestation land use classification.

We used such a simple encoding since we did not obtain better results with other encoder architectures we tested—a deeper encoder consisting of multiple FC layers, and a separate linear projection layers for each tabular value.

4.4. Multi-modal fusion and classification

After encoding input modalities with individual encoders, we obtain a set T of t tokens, each of size z. To fuse the tokens together and extract a multi-modal representation, we used the Transformer encoder (Vaswani et al., 2017), $Enc_{\text{transf}}(T, \theta_{transf}) : \mathbb{R}^{t \times z} \rightarrow \mathbb{R}^{t \times z}$ (middle part of Fig. 4). The Transformer requires all tokens to be of a fixed size, which was the motivation for using modality-specific encoders. The self-attention layers utilized by the Transformer allow each token to attend to all other input tokens, enabling interactions across different input modalities.

The encoder output is of the same dimensionality as its input, i.e., T tokens of dimension z. To adapt it for classification purposes, we add one randomly initialized 'CLS' (classification) token at the beginning of the input token sequence, as proposed by Devlin (2018). The output embedding corresponding to the 'CLS' token is an aggregated representation of the entire multi-modal token sequence. This is the standard approach for adapting the Transformer architecture for the classification task both in language and vision (Devlin, 2018; Dosovitskiy et al., 2020). Other options exist, such as a global average pooling of all output tokens, however no benefit of such an approach has been observed (Dosovitskiy et al., 2020).

This 'CLS' token is then passed to a FC layer ('Classifier' in Fig. 4) $FC(T, \theta_{FC}) : \mathbb{R}^z \rightarrow \mathbb{R}^{11}$ with a soft-max activation function to produce probabilities for the 11 output classes.

5. Experimental setup

The experiments were carried out to test our hypothesis that including location and tabular data modalities would improve the classification accuracy compared to using only image time series. This was analyzed with different sizes of the training dataset because the scarcity of labeled training data is one of the major challenges associated with classifying post-deforestation land use.

Additional experiments were conducted to:

- evaluate the proposed multi-modal fusion method against the following baselines
  - mid-fusion with encoders: tokens are extracted from all modalities with modality-specific encoders and concatenated; the resulting vector is passed to a FC classifier (Fig. B.9(a))
  - mid-fusion (approach of Irvin et al. 2020): only image encoder is used; image tokens are concatenated with other inputs and the resulting feature vector is passed to a FC classifier (Fig. B.9(b))
  - late fusion (approach of Masolele et al. 2022): tokens are extracted from all modalities with modality-specific encoders, followed by individual FC classifiers for each modality; modality-specific predictions are averaged (Fig. B.9(c))
- test the hypothesis that using the deforestation mask as an input will improve the model's performance;
- compare different methods of location encoding and identify which PE and NN achieves the best performance;
- find the optimal size of the image time series;

Finally, the location and tabular modalities were each analyzed to provide insights and intuition of the behavior of the model.

The model was trained end-to-end by minimizing the cross-entropy loss:

$$-\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log(p_{i,c}) \tag{1}$$

where N is the number of training samples, C is the number of classes, $y_{i,c} \in \{0,1\}$ is the ground-truth label for sample i, and $p_{i,c}$ is the predicted probability obtained via a softmax over the model logits.

For the ResNet18 image encoder we used weights pre-trained on ImageNet. The pre-trained weights only have three input channels as they were trained on natural RGB images, while we use four bands. To account for this, we duplicated the weights corresponding to the first channel to match the shape of our input data.

All experiments were executed on a single NVIDIA GeForce RTX 3090 GPU. We tuned all hyperparameters using 5-fold cross validation. To obtain the final test metrics, we trained the model for 50 epochs. All training configurations were executed 5 times with different random seeds and the results were averaged.

As evaluation metrics, we use accuracy and F1 score.

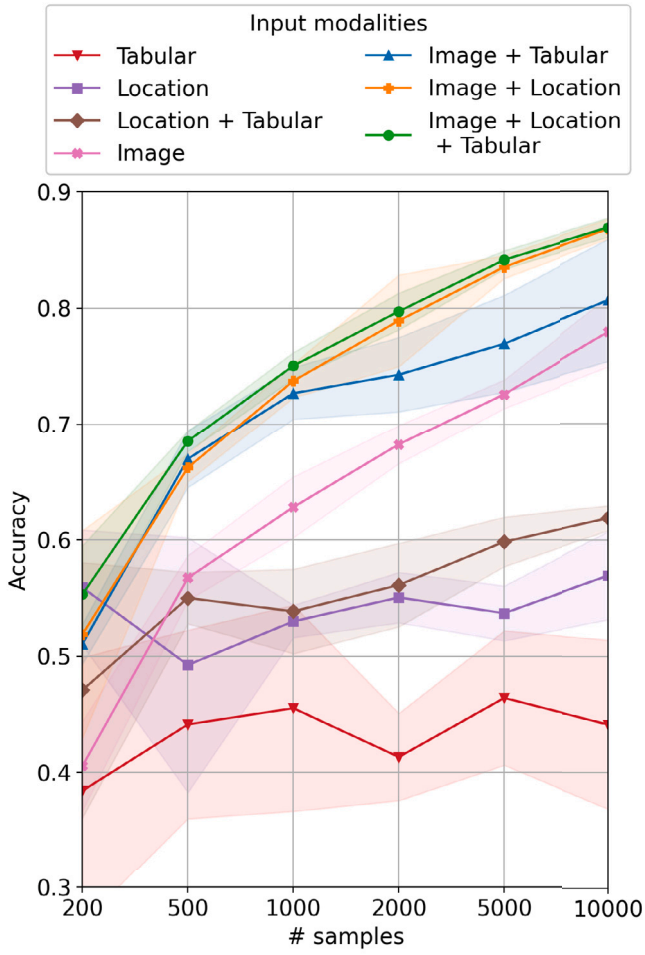


Fig. 5. Accuracy for different combinations of input modalities at different training dataset sizes.

6. Results

We first report and discuss the impact of using different combinations of input data modalities. Then, we report on ablation studies and sensitivity analyses. Finally, we analyze the location and tabular modalities.

6.1. Input modality combinations

Figure 5 shows the results of all combinations of input modalities for varying training dataset sizes. The best results are obtained when using all three modalities, which is notable mostly in low-data regimes. When training on the full dataset size, the model trained on all modalities is on par with the model using images and location only.

With increasing training dataset size, the inclusion of location brings larger improvements than tabular data. This is likely due to the nature of the tabular data, which is only available at a country level, and therefore increasing the dataset size has only limited effect (the number of countries remains the same). In contrast, every training sample is associated with a unique pair of coordinates, corresponding to more consistent improvements with increasing the training dataset size. We also find that the benefit of tabular data depends on country size. As shown in Fig. E.12, the improvements in F1 score decrease with increasing country area, because larger countries have more internal variability that is not well captured by a single set of values.

Table 2
Per-Class F1-Scores for Different Model Configurations.

Input modality	**Model configuration**			
Image	x	x	x	x
Location			x	x
Tabular		x		x
Class name	**F1-Score**			
Buildings/roads	**0.93**	**0.93**	0.92	**0.93**
Cocoa	0.90	0.88	**0.96**	0.95
Coffee	0.44	**0.63**	0.61	0.60
Forest Plantation	0.65	**0.66**	**0.66**	0.65
Mining	**0.86**	0.84	0.85	0.84
Oil Palm	0.68	0.69	0.80	**0.82**
Other	0.77	0.83	**0.87**	**0.87**
Pasture	0.76	0.85	**0.86**	0.85
Rubber	0.58	0.66	0.87	**0.88**
Shifting Agriculture	0.73	0.83	0.95	**0.97**
Soy	**0.85**	**0.85**	**0.85**	**0.85**
Weighted Average	0.78	0.81	**0.87**	**0.87**
Macro Average	0.74	0.79	**0.84**	**0.84**

Satisfactory results are achieved using only the image modality, with consistent improvements when increasing the training dataset. This is as expected because including more image samples increases the diversity of the training data. Nevertheless, there remains a consistent gap of approximately 10% accuracy between the image-only and the multi-modal model. For all variants including the image modality, the shape of the plotted lines suggests that further improvements could be gained if more data was available for training.

When using only the location as input, training with a smaller dataset results in high variance, suggesting instability and lack of robustness. With increasing the training dataset size, the accuracy also increases and reaches 0.57 at full dataset size, confirming the presence of strong spatial patterns in the data.

Using the country-level tabular data as the only input is clearly limited, however modest gains in accuracy can be observed when increasing the training dataset up to 0.44 accuracy when trained on the full dataset size. This is because samples within the same country are not always identical, since each training sample can be associated with different deforestation years. This introduces slight variations to the input data, as they are computed with respect to the deforestation period of each input sample.

Training on the combination of the two auxiliary modalities results in better performance (0.62) than using each individually, suggesting they contain complementary information. Nevertheless, the image time series is clearly the most important input modality. The task of land use classification cannot be performed accurately without visual cues. Therefore, we only consider modality combinations that include the image time series in the rest of this section.

Per-class results for the full dataset size are shown in Table 2. For most classes, the multi-modal variants outperform the image-only baseline. For the recognition of classes *buildings/roads*, *mining* and *soy*, the additional modalities bring no benefits. The former two are generally not tied to specific locations or regions, and are not associated with tabular data (which is limited to agricultural commodities). In contrast, tabular data is available for *soy*, which is also clearly spatially clustered — all examples of this class in the dataset are found in three South American countries, Brazil, Uruguay and Paraguay. Nevertheless, all three classes have a distinct visual appearance (shown in Fig. 1), allowing accurate classification from satellite images only.

Poorest results are obtained on classes *coffee* and *forest plantation*. With respect to *coffee*, there were fewer training examples available for this class. Also, there are multiple cultivation styles of coffee, such as sun-grown and shade-grown (Pisl et al., 2024a). These are associated

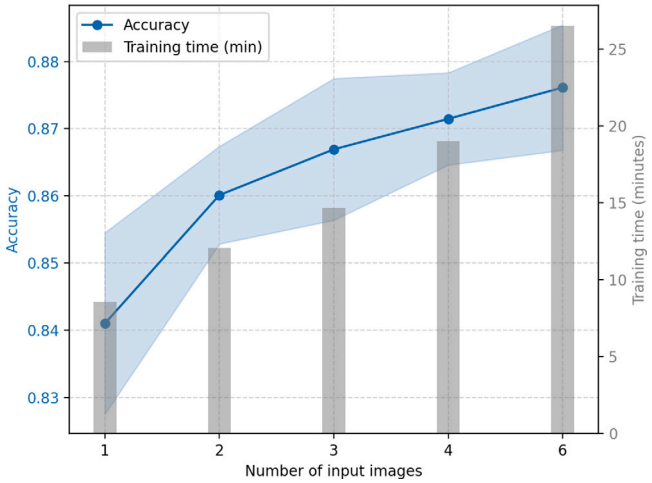


Fig. 6. Accuracy for different lengths of the input time series and corresponding training times. The time series is always sampled from the year 2023.

with different appearance on satellite imagery, causing high intra-class variability, This makes the recognition of this class particularly difficult.

As for *forest plantation*, it likely is not associated with strong enough visual or location patterns. Since the labeled examples of this class come from only a single dataset, it is also possible there is a higher rate of label noise. As for the lack of benefits from using tabular data, this may be explained by the fact that the commodity associated with this class (wood fiber) is also produced by extracting wood from natural forests (as opposed to forest plantations). Therefore, if the tabular data indicate high deforestation rates due to wood extraction, it does not have to correspond to areas being converted to forest plantations. This is in contrast to all the other included commodities.

6.2. Ablation & sensitivity analysis

Multi-modal fusion. The results are shown in Table 3. The Transformer encoder achieves the best results, outperforming the other mid-fusion variants by 2% and 3%, respectively. A greater drop of 5% in performance is observed when using late fusion, suggesting cross-modal interactions are important for this task. As for the model size, the Transformer encoder is larger than MLPs, however its size accounts only for approximately 14% of the total model size, since the largest part of the model is the image encoder (11 million parameters).

Deforestation mask. Using the deforestation mask in the image encoder results in an improvement of 5% as shown in Table 4. Without the deforestation mask, the model receives no indication of which part of the image is relevant, and extracts features from the entire image. In many cases, this can lead to a correct prediction since the images covers a relatively small area of 1 km^2, and a single land use may dominate the entire area. However, as the results demonstrate, often the precise localization of the deforestation within the image is necessary.

Comparison of location encoders. The differences between the different variants of location encoding are minimal as shown in Table 5. We use the combination of "Wrap" as a PE method and "Linear" as a NN due to the joint best performance and lowest number of parameters. Due to the small differences in obtained results, we performed a pair-wise t-test which confirmed this selection (Fig. C.10).

Table 3
Comparison of multi-modal fusion; the architectures are shown in Fig. B.9.

Fusion type	Architecture	# of parameters	Accuracy
mid-fusion	encoders & MLP	11 636 300	0.84
mid-fusion	MLP	11 426 884	0.85
late fusion	MLP per modality	11 442 786	0.82
mid-fusion	encoders & Transformer (ours)	13 028 108	**0.87**

Table 4
Usage of deforestation mask in image encoder.

Deforestation mask	Accuracy
No	0.82
Yes	**0.87**

Image time series length. The results for varying number of input images are shown in Fig. 6, demonstrating an expected trade-off where increasing the number of input images improves the accuracy, but also increases training time. The training times are low since the largest part of the model – the image encoder – is pre-trained and the rest is relatively small. However, considering the decrease in improvements from using longer time series, and especially the amount of data needed to train the model, we choose to use a time series consisting of four images.

6.3. Analysis of the location modality

We focus on two classes – *buildings/roads* and *oil palm* – to investigate how including the location modality impacts the prediction of the model. As discussed in Section 6.1, the class *buildings/roads* can be accurately recognized from satellite images alone and is not associated with strong spatial clustering-deforestation followed by establishing roads or buildings is a common scenario anywhere in the tropics. Therefore, we hypothesize that this class can be recognized by the model regardless of the spatial location.

In contrast, *oil palm* is a tree crop with arguably less distinct visual appearance, and spectral signature similar to other tree crop classes, such as *coffee*, *cocoa*, *rubber*, or *forest plantation*. It is also much more likely to appear in certain regions, particularly in Malaysia and Indonesia. The results in Table 2 confirm that including the location modality indeed significantly improves the model's ability to accurately classify *oil palm*. We expect the model to be more sensitive to changes in location with respect to this class.

To test our hypotheses, we generate 1000 random locations across the tropics, and we randomly select one image time series of the *oil palm* class. We use the trained model to repeatedly classify the same image time series 1000 times, each time paired with a different location. We then perform the same experiment with an image time series showing *buildings/roads*. To isolate the impact of the location modality, we do not use tabular data in this experiment. This allows us to observe how changing the input coordinates impacts the model's prediction for each of the two classes, while keeping other inputs fixed.

The results confirm our hypotheses. For the class *buildings/roads*, the model correctly predicts the class in all 1000 cases. This suggests that the model can recognize this class from the input images and does not rely on the location modality. In contrast, the input coordinates play a significant role when classifying the image time series of *oil palm*. Out of the 1000 samples, the model predicts 719 samples as *oil palm*, 55 as *rubber*, and 226 as *forest plantation*. The predictions are clearly spatially clustered as shown in Fig. 7. All of the *rubber* predictions are located in Southeast Asia, around Myanmar, Thailand, Laos, Cambodia and Vietnam. This corresponds to the spatial distribution of the training

Table 5
Comparison of location encoder variants.

PE	NN	# of parameters	Accuracy
Direct	Siren	33 408	0.86
Direct	Linear	384	0.86
Direct	FCnet	132 480	0.86
Wrap	Siren	33 664	0.86
Wrap	Linear	640	**0.87**
Wrap	FCnet	132 864	**0.87**
Cartesian3D	Siren	33 536	**0.87**
Cartesian3D	Linear	512	0.86
Cartesian3D	FCnet	132 736	0.86

data, where *rubber* is the most common class in that region as shown in Fig. 2. All of the locations where the model predicts *forest plantation* are located in the central part of South America in Brazil, Bolivia, Peru and Chile. While there are many training samples of *forest plantation* in that region, in a large part of the region, the most common class is *soy* (Fig. 2). Despite that, the model does not predict any *soy* in this experiment, likely because *soy* has a distinctly different visual appearance.

These results suggest that, when the image data is insufficient for the recognition of the ground truth class, the model does not simply revert to predicting the most common class in a given region. Instead, it chooses from other plausible options, in this case *forest plantation*, which has a similar appearance to *oil palm*.

6.4. Analysis of the tabular modality

Here, we investigate how the tabular data impacts the prediction. The tabular data consists of a set of eleven values, one per output class. They are coarse estimates of how much deforestation can be attributed to a given class (or a placeholder in case of missing values). Intuitively, a higher value for a certain class should therefore increase the output probability of a given class. The link between the input tabular values and the output classes is not made explicit in the model architecture, but we hypothesize that the model learned this implicitly during training. In this experiment, we aim to confirm or reject this hypothesis.

We use the test samples within the country of Cambodia because all of its samples lie in the test set, therefore avoiding the possibility of a leakage of the tabular data between the train and test sets. We select three classes, *rubber* as an example of an agricultural commodity and the predominant class in the country, *soy* as another commodity which however is not present in the country at all, and *mining* as a land use class not related to agriculture and therefore not associated with tabular values. We make repeated predictions on all samples of each class, each time with a modified tabular data value for a given class. Instead of using the tabular value in our dataset, we test values between of 0 and 3 with increments of 0.1. The selected range reflects the fact that the tabular values have been normalized to 0 mean and unit variance, and therefore 99.7% of the values will lie between 0 and three standard deviations, i.e., 3. We observe how the output probabilities of the model change as a result of the modified tabular values. If the model learns to implicitly associate the tabular values with the output classes they correspond to, the average predicted probability for a class should increase as the tabular value increases.

The results are shown in Fig. 8 and confirm our hypothesis. When increasing the tabular values for the class *rubber*, it generally indeed corresponds to an increase of the average probability prediction of this class, with the exception of tabular values in range between 0 and 0.1. The same trend can be seen when augmenting the tabular value for *soy*, although the shape of the curve is different. In contrast, with *mining* this cannot be observed. The modifications of the tabular value have an irregular impact on the model predictions. This can be explained by

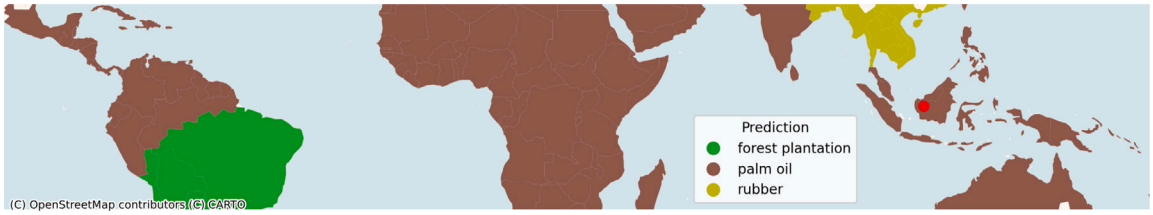
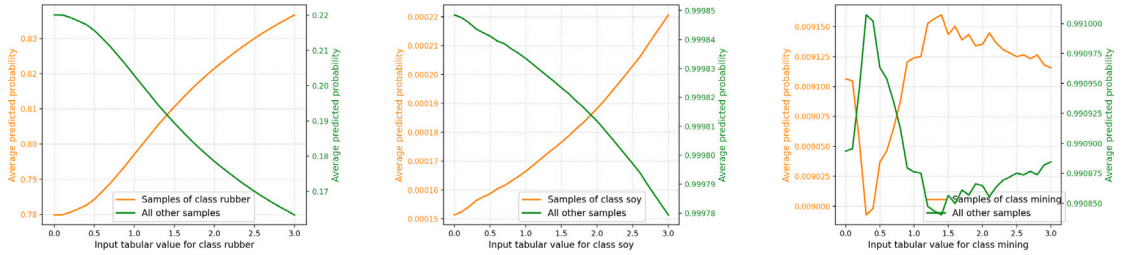


Fig. 7. Prediction of the same fixed time series of an oil palm plantation, with varying randomly generated locations; Voronoi polygons created from the predictions for visualization; the original location of the time series is indicated with a red circle. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



(a) Rubber (b) Soy (c) Mining

Fig. 8. Tabular data experiment — average class probabilities on samples in the test country of Cambodia as a function of input tabular values.

the fact that no tabular data was available for this class during training (i.e., it was always replaced by the missing data indicator), therefore the model did not learn the association between this particular tabular value and the output class.

7. Discussion

In this work, we propose a new approach for mapping post-deforestation land use that combines satellite image time series with two types of location inputs. We present a Transformer-based model architecture designed specifically for this task which integrates all input data in a flexible and efficient manner. We demonstrate that adding the geographic location and coarse, country-level tabular data improves the resulting accuracy, the latter in particular when training on smaller datasets. The use of the precise location of the deforestation events within the image, provided by a deforestation map, results in further performance gains. The results of our experiments show while the choice of model architecture has an impact on the results, the selection of input data is a more important factor.

We train the model on a pan-tropical dataset, which we compile from various public sources to enable the recognition of the land uses associated with deforestation from freely available data. This can contribute to closing the spatial and temporal gap in the knowledge of the deforestation drivers, since post-deforestation land use is a common proxy for the estimation of drivers. We show that the size of the training dataset has a strong impact on the model's performance and collecting more data would likely bring additional gains.

It is important to point out some limitations of the presented approach. The tabular data bring modest gains in performance that decrease with increasing the training dataset size. We believe this is due to the coarse spatial resolution and subsequently insufficient amount of unique training data points to learn from. As discussed below, this may be addressed by new, more detailed datasets. To some extent, this is likely also due to the imprecision of the country-level statistics currently available which in some cases fundamentally differ from results obtained through remote sensing (Kalischek et al., 2023).

As for the location modality, the model can only learn spatial patterns that are present in the training data. As is the nature of neural networks, our model is not capable to extrapolate to new, unseen regions. Since the training data is spatially biased, this bias can be transferred into the model during the training process. It is therefore possible that the model might not reach the reported performance on some classes in certain geographical regions where training samples were not available. This highlights the need for more, and more evenly distributed training data.

There are several promising directions for future work. Because of the flexible, Transformer-based architecture, more input modalities can be added as inputs. Specifically, environmental, climatic and topographic variables can be valuable covariates because agricultural commodities can typically be cultivated only in specific conditions, or are associated with low (e.g., oil palm) or high (e.g., coffee) altitude. Data from other sensors, such as Sentinel-1, can also be integrated, to tackle the high cloud cover often present in the tropics.

The country-level tabular data on agricultural commodities could be replaced with higher resolution alternatives as they start to become available. Goldman et al. (2020) provides estimates of the post-deforestation land use at a county level for the seven forest-risk commodities across tropics. The CROPGRIDS dataset (Tang et al., 2024) provides the crop and harvest area for 173 crops globally at a 10 km resolution. Such data can be integrated in a similar way to the DeDuCE dataset but might yield significant improvements since it holds much more fine-grained information about the spatial distribution of the commodities.

Overall, this work proposes a new, data-driven approach for the fine-grained classification of post-deforestation land use, using recent advancements in multi-modal deep learning to overcome current limitations. It can be used to attribute deforestation anywhere in the tropics to specific drivers in an automatic and repeatable fashion. Understanding of the drivers is crucial for designing and implementing effective and targeted policy responses. As such, we hope our work contributes to the protection and preservation of the tropical forests.

CRediT authorship contribution statement

Jan Pišl: Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Gencer Sumbul:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Gaston Lenczner:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Camilo Zamora:** Writing – review & editing, Data curation. **Martin Herold:** Writing – review & editing. **Jan Dirk Wegner:** Conceptualization. **Devis Tuia:** Writing – review & editing, Supervision, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 945363.

Appendix A. Location encoder variants

A.1. Positional embedding

We compare the following three positional embedding methods, denoting latitude and longitude as λ, φ:

- **Direct (no embedding)**

$$\mathrm{PE}(\lambda, \varphi) = (\lambda, \varphi)$$

- **Wrap** (Mac Aodha et al., 2019)

$$\mathrm{PE}(\lambda, \varphi) = \left[\cos\lambda, \quad \sin\lambda, \quad \cos\varphi, \quad \sin\varphi\right]$$

- **3D Cartesian** (Tseng et al., 2022)

$$\mathrm{PE}(\lambda, \varphi) = \left[\cos(\mathrm{lat}) \times \cos(\mathrm{lon}), \quad \cos(\mathrm{lat}) \times \sin(\mathrm{lon}), \quad \sin(\mathrm{lat})\right]$$

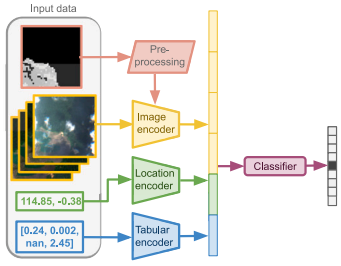
A.2. Neural network

We compare the following three encoders:

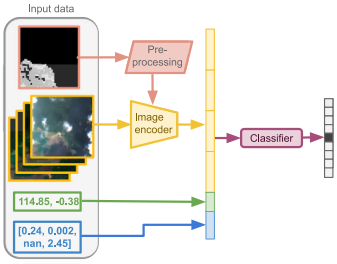
- **Linear** corresponds to a single linear layer with no activation function which transforms the output of the PE to a token of size z.
- **FCNet** (Mac Aodha et al., 2019) is a FC architecture proposed specifically as a location encoder. It consists of an input linear layer, four residual modules and an output layer. Each residual module contains two FC layers with ReLU activations, a dropout, and a residual connection.
- **SIREN** (SInusoidal REpresentation Network) (Sitzmann et al., 2020) is an architecture that replaces standard activation functions (e.g., ReLU) with a sinusoidal function. The implementation used in this work contains two FC layers, each followed by a dropout and a sinusoidal activation function and a final linear layer without activation.

Appendix B. Multi-modal fusion baselines

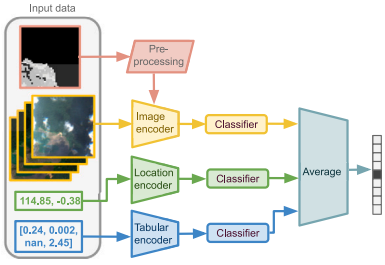
See Fig. B.9.



(a) Mid-fusion with modality-specific encoders; extracted tokens are concatenated together and the resulting feature vector is passed to a FC classifier



(b) Mid-fusion with image encoder only (approach of (Irvin et al., 2020)); image tokens are concatenated with other inputs and the resulting feature vector is passed to a FC classifier



(c) Late fusion (approach of (Masolele et al., 2022)); modality-specific encoders are followed by individual classifiers for each modality; the predictions are then averaged

Fig. B.9. Multi-modal fusion baselines.

Appendix C. Location encoder results — statistical significance test

See Fig. C.10.

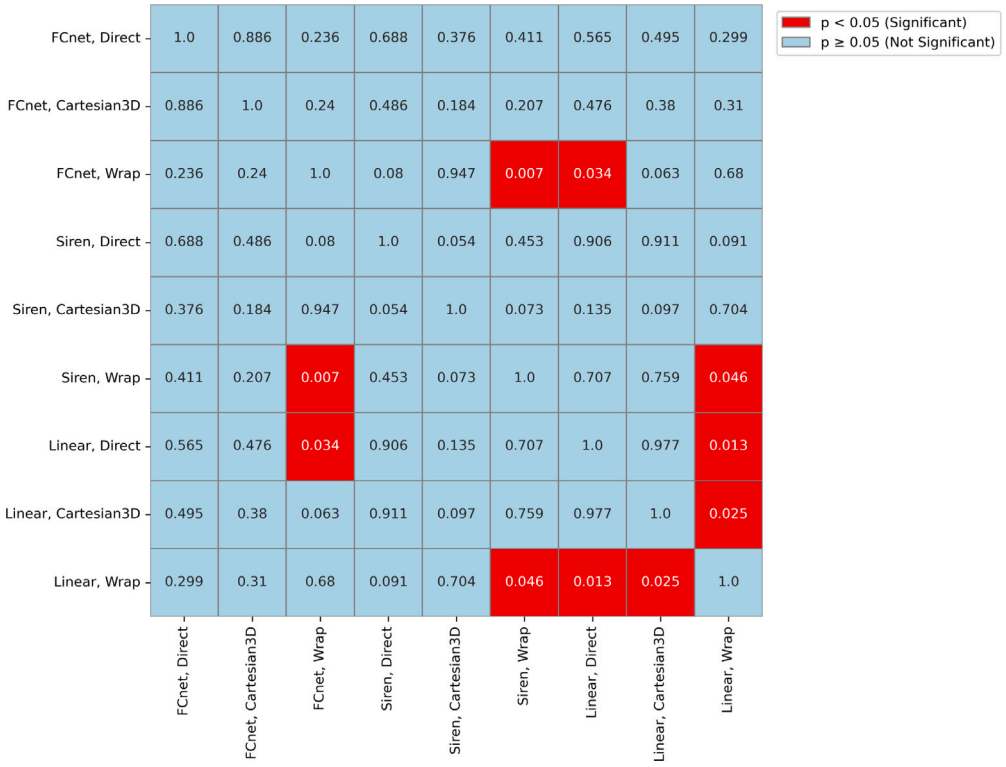


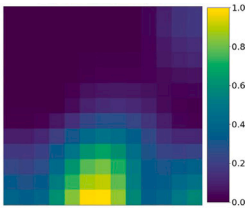
Fig. C.10. Pair-wise t-test of statistical significance of the difference between accuracies achieved with the tested location encoder variants; the differences between most variants are not statistically significant, therefore we choose the combination of "Wrap" as a PE method and "Linear" as a NN; this is the most efficient variant (in terms of number of parameters) not outperformed by any other.

Appendix D. Prediction examples

See Fig. D.11.

a)

b)



c)

d)

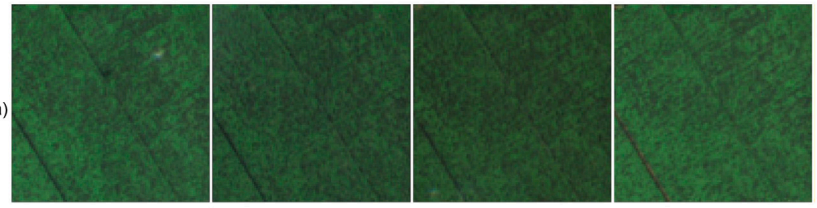
Class name	pasture	forest plantation	soy	coffee	palm oil	mining	urban	cocoa	rubber	shifting agriculture	other
Tabular value	0.02	0.23	*no data*	0.03	0.78	*no data*	*no data*	0.03	*no data*	*no data*	*no data*

e)

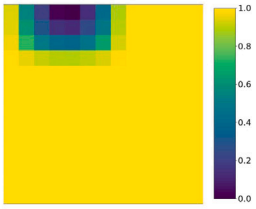
Class name	pasture	forest plantation	soy	coffee	palm oil	mining	urban	cocoa	rubber	shifting agriculture	other
Model prediction	0	0	0	0	**0.51**	0.02	0	0.44	0.01	0.02	0

(a) Correctly predicted sample

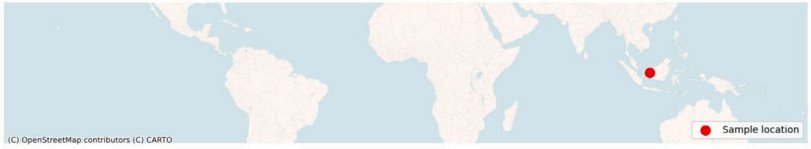
a)



b)



c)



d)

Class name	pasture	forest plantation	soy	coffee	palm oil	mining	urban	cocoa	rubber	shifting agriculture	other
Tabular value	0.04	0.07	0	0.01	3.49	*no data*	*no data*	0	0.53	*no data*	*no data*

e)

Class name	pasture	forest plantation	soy	coffee	palm oil	mining	urban	cocoa	rubber	shifting agriculture	other
Model prediction	0	0.28	0	0	**0.37**	0.01	0	0	0.33	0	0

(b) Incorrectly predicted sample

Fig. D.11. Examples of two test samples and the corresponding predictions made by the model; (a) satellite image time series; (b) smoothed deforestation mask (value of 1 indicates deforestation, 0 indicates no deforestation); (c) geographic location of the sample visualized on a map; (d) input tabular data (e) class probabilities produced by the model (highest probability is **bold**, correct class is underscored).

Appendix E. Influence of country size on performance improvement of using tabular data

See Fig. E.12.

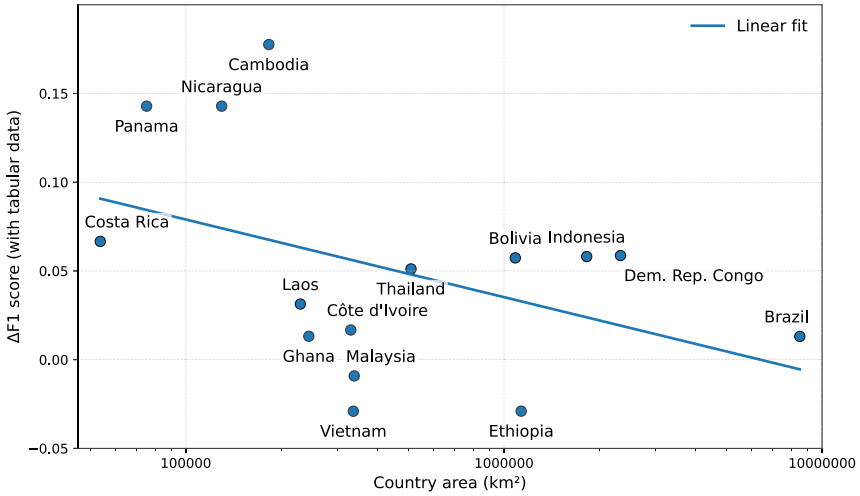


Fig. E.12. Plot indicating that smaller countries benefit more from the inclusion of tabular data; We computed the difference in test F1 score obtained by models with and without using the tabular data for all countries present in the test set; the difference was computed using models trained on all training dataset sizes and averaged.

References

Aleissaee, A.A., Kumar, A., Anwer, R.M., Khan, S., Cholakkal, H., Xia, G.-S., Khan, F.S., 2023. Transformers in remote sensing: A survey. Remote. Sens. 15 (7), 1860.

Alonso, L., Picos, J., Armesto, J., 2022. Automatic identification of forest disturbance drivers based on their geometric pattern in Atlantic forests. Remote. Sens. 14 (3), 697.

Becerra, M., Rivera, O., Pinto, N., 2022. Base de datos de cobertura de cultivos de cacao en la Amazonia Peruana. http://dx.doi.org/10.7910/DVN/XMQNC2, Dataset.

Bernhard, K.P., Shapiro, A.C., Hunt, C.A., 2024. Drivers of tropical deforestation: a global review of methodological approaches and analytical scales. Biodivers. Conserv. 33 (1), 1–29.

Brown, J.H., 2014. Why are there so many species in the tropics? J. Biogeogr. 41 (1), 8–22.

Campos-Taberner, M., García-Haro, F.J., Martínez, B., Izquierdo-Verdiguier, E., Atzberger, C., Camps-Valls, G., Gilabert, M.A., 2020. Understanding deep learning in land use classification based on sentinel-2 time series. Sci. Rep. 10 (1), 17188.

Ceccarelli, V., Fremout, T., Zavaleta, D., Lastra, S., Imán Correa, S., Arévalo-Gardini, E., Rodriguez, C.A., Cruz Hilacondo, W., Thomas, E., 2021. Climate change impact on cultivated and wild cacao in Peru and the search of climate change-tolerant genotypes. Diversity and Distributions 27 (8), 1462–1476.

Curtis, P.G., Slay, C.M., Harris, N.L., Tyukavina, A., Hansen, M.C., 2018. Classifying drivers of global forest loss. Science 361 (6407), 1108–1111.

De Sy, V., Herold, M., Achard, F., Avitabile, V., Baccini, A., Carter, S., Clevers, J.G., Lindquist, E., Pereira, M., Verchot, L., 2019. Tropical deforestation drivers and associated carbon emission factors derived from remote sensing data. Environ. Res. Lett. 14 (9), 094022.

Descals, A., Wich, S., Meijaard, E., Gaveau, D.L., Peedell, S., Szantoi, Z., 2020. High-resolution global map of smallholder and industrial closed-canopy oil palm plantations. Earth Syst. Sci. Data Discuss. 2020, 1–22.

Devlin, J., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16 × 16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

FAO, 2020. Global Forest Resources Assessment. FAO, http://dx.doi.org/10.4060/ca9825en.

Food and Agriculture Organization of the United Nations (FAO), 2025. FAOSTAT. https://www.fao.org/faostat/en/#data. (Accessed 13 June 2025).

Fritz, S., Laso Bayas, J.C., See, L., Schepaschenko, D., Hofhansl, F., Jung, M., Dürauer, M., Georgieva, I., Danylo, O., Lesiv, M., McCallum, I., 2022. A continental assessment of the drivers of tropical deforestation with a focus on protected areas. Front. Conserv. Sci. 3, 830248.

Garnot, V.S.F., Landrieu, L., Chehata, N., 2022. Multi-modal temporal attention models for crop mapping from satellite time series. ISPRS J. Photogramm. Remote Sens. 187, 294–305.

Geist, H.J., Lambin, E.F., 2002. Proximate causes and underlying driving forces of tropical deforestation. BioScience 52 (2), 143.

Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I., 2023. Imagebind: One embedding space to bind them all. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15180–15190.

Goldman, E., Weisse, M., Harris, N., Schneider, M., 2020. Estimating the Role of Seven Commodities in Agriculture-Linked Deforestation: Oil Palm, Soy, Cattle, Wood Fiber, Cocoa, Coffee, and Rubber. World Resources Institute.

Hansen, M.C., Krylov, A., Tyukavina, A., Potapov, P.V., Turubanova, S., Zutta, B., Ifo, S., Margono, B., Stolle, F., Moore, R., 2016. Humid tropical forest disturbance alerts using Landsat data. Environ. Res. Lett. 11 (3), 034008.

Hansen, M., Potapov, P., Margono, B., Stehman, S., Turubanova, S., Tyukavina, A., 2014. Response to comment on "high-resolution global maps of 21st-century forest cover change". Science 344 (6187).

Hansen, M.C., Potapov, P.V., Moore, R., Hancher, M., Turubanova, S.A., Tyukavina, A., Thau, D., Stehman, S.V., Goetz, S.J., Loveland, T.R., et al., 2013. High-resolution global maps of 21st-century forest cover change. Science 342 (6160), 850–853.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

Hermosilla, T., Wulder, M.A., White, J.C., Coops, N.C., Hobart, G.W., 2015. Regional detection, characterization, and attribution of annual forest change from 1984 to 2012 using Landsat-derived time-series metrics. Remote Sens. Environ. 170, 121–132.

Hosonuma, N., Herold, M., De Sy, V., De Fries, R.S., Brockhaus, M., Verchot, L., Angelsen, A., Romijn, E., 2012. An assessment of deforestation and forest degradation drivers in developing countries. Environ. Res. Lett. 7 (4), 044009.

Irvin, J.A., Sheng, H., Ramachandran, N., Johnson-Yu, S., Zhou, S., Story, K., Rustowicz, R., Elsworth, C., Austin, K., Ng, A., 2020. ForestNet: Classifying drivers of deforestation in Indonesia using deep learning on satellite imagery. In: NeurIPS Workshop on Tackling Climate Change with Machine Learning.

Kalischek, N., Lang, N., Renier, C., Daudt, R.C., Addoah, T., Thompson, W., Blaser-Hart, W.J., Garrett, R., Schindler, K., Wegner, J.D., 2022. Satellite-based high-resolution maps of cocoa planted area for Côte d'Ivoire and Ghana.

Kalischek, N., Lang, N., Renier, C., Daudt, R.C., Addoah, T., Thompson, W., Blaser-Hart, W.J., Garrett, R., Schindler, K., Wegner, J.D., 2023. Cocoa plantations are associated with deforestation in Cote d'Ivoire and Ghana. Nat. Food 4 (5), 384–393.

Kaselimi, M., Voulodimos, A., Daskalopoulos, I., Doulamis, N., Doulamis, A., 2022. A vision transformer model for convolution-free multilabel classification of satellite imagery in deforestation monitoring. IEEE Trans. Neural Networks Learn. Syst. 34 (7), 3299–3307.

Kissinger, G., Herold, M., De Sy, V., 2012. Drivers of Deforestation and Forest Degradation: A Synthesis Report for REDD+ Policymakers. Technical Report, Lexeme Consulting.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. 25.

Laso Bayas, J.C., See, L., Georgieva, I., Schepaschenko, D., Danylo, O., Dürauer, M., Bartl, H., Hofhansl, F., Zadorozhniuk, R., Burianchuk, M., et al., 2022. Drivers of tropical forest loss between 2008 and 2019. Sci. Data 9 (1), 146.

Lewis, S.L., Edwards, D.P., Galbraith, D., 2015. Increasing human dominance of tropical forests. Science 349 (6250), 827–832.
Li, J., Hong, D., Gao, L., Yao, J., Zheng, K., Zhang, B., Chanussot, J., 2022. Deep learning in multimodal remote sensing data fusion: A comprehensive review. Int. J. Appl. Earth Obs. Geoinf. 112, 102926.
Liu, Q., Kampffmeyer, M., Jenssen, R., Salberg, A.-B., 2022. Multi-modal land cover mapping of remote sensing images using pyramid attention and gated fusion networks. Int. J. Remote Sens. 43 (9), 3509–3535.
Mac Aodha, O., Cole, E., Perona, P., 2019. Presence-only geographical priors for fine-grained image classification. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 9596–9606.
MapBiomas Bolivia, 2024. MapBiomas project - Collection 2021 of the annual land use land cover maps of Bolivia.
MapBiomas Brazil, 2024. MapBiomas project - Collection 2022 of the annual land use land cover maps of Brazil.
MapBiomas Indonesia, 2024. MapBiomas project - Collection 2022 of the annual land use land cover maps of Indonesia.
MapBiomas Paraguay, 2024. MapBiomas project - Collection 2022 of the annual land use land cover maps of Paraguay.
MapBiomas Peru, 2024. MapBiomas project - Collection 2022 of the annual land use land cover maps of Peru.
Maskell, G., Chemura, A., Nguyen, H., Gornott, C., Mondal, P., 2021. Integration of sentinel optical and radar data for mapping smallholder coffee production systems in Vietnam. Remote Sens. Environ. 266, 112709.
Masolele, R.N., De Sy, V., Herold, M., Marcos Gonzalez, D., Verbesselt, J., Gieseke, F., Mullissa, A.G., Martius, C., 2021. Spatial and temporal deep learning methods for deriving land-use following deforestation: A pan-tropical case study using Landsat time series. Remote Sens. Environ. 264, 112600, Publisher: Elsevier.
Masolele, R.N., De Sy, V., Marcos, D., Verbesselt, J., Gieseke, F., Mulatu, K.A., Moges, Y., Sebrala, H., Martius, C., Herold, M., 2022. Using high-resolution imagery and deep learning to classify land-use following deforestation: a case study in Ethiopia. GIScience & Remote. Sens. 59 (1), 1446–1472.
Masolele, R.N., Marcos, D., De Sy, V., Abu, I.-O., Verbesselt, J., Reiche, J., Herold, M., 2024. Mapping the diversity of land uses following deforestation across Africa. Sci. Rep. 14 (1), 1681.
Maus, V., da Silva, D., Gutschlhofer, J., da Rosa, R., Giljum, S., Gass, S.L.B., Luckeneder, S., Lieber, M., McCallum, I., 2022. Global-scale mining polygons (version 2).
Mena, F., Arenas, D., Nuske, M., Dengel, A., 2024. Common practices and taxonomy in deep multiview fusion for remote sensing applications. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 17, 4797–4818.
Mitton, J., Murray-Smith, R., 2021. Rotation equivariant deforestation segmentation and driver classification. In: NeurIPS Workshop on Tackling Climate Change with Machine Learning.
Mullissa, A., Reiche, J., Saatchi, S., 2023. Seasonal forest disturbance detection using sentinel-1 SAR & sentinel-2 optical timeseries data and transformers. In: IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium. pp. 3122–3124.
Nguyen, T.H., Jones, S.D., Soto-Berelov, M., Haywood, A., Hislop, S., 2018. A spatial and temporal analysis of forest dynamics using Landsat time-series. Remote Sens. Environ. 217, 461–475.
Pastorino, M., Gallo, F., Di Febbraro, A., Moser, G., Sacco, N., Serpico, S.B., 2022. Multimodal fusion of mobility demand data and remote sensing imagery for urban land-use and land-cover mapping. Remote. Sens. 14 (14), 3370.
Pendrill, F., Gardner, T.A., Meyfroidt, P., Persson, U.M., Adams, J., Azevedo, T., Bastos Lima, M.G., Baumann, M., Curtis, P.G., De Sy, V., et al., 2022. Disentangling the numbers behind agriculture-driven tropical deforestation. Science 377.
Peng, D., Cheng, E., Feng, X., Hu, J., Lou, Z., Zhang, H., Zhao, B., Lv, Y., Peng, H., Zhang, B., 2024. A deep–learning network for wheat yield prediction combining weather forecasts and remote sensing data. Remote. Sens. 16 (19), 3613.
Petersen, R., Goldman, E.D., Harris, N., Sargent, S., Aksenov, D., Manisha, A., Esipova, E., Shevade, V., Loboda, T., Kuksina, N., et al., 2016. Mapping tree plantations with multispectral imagery: preliminary results for seven tropical countries. World Resour. Inst. Wash. DC 525.
Pisl, J., Lenczner, G., Tuia, D., De Morsier, F., 2024a. Semantic segmentation of coffee plantations from sentinel-2 time series. In: IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium. IEEE, pp. 1526–1530.
Pisl, J., Rußwurm, M., Hughes, L., Lenczner, G., See, L., Wegner, J.D., Tuia, D., 2024b. Mapping drivers of tropical forest loss with satellite image time series and machine learning. Environ. Res. Lett. 6.
Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., 2021. Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. PmLR, pp. 8748–8763.
Richards, D.R., Friess, D.A., 2016. Rates and drivers of mangrove deforestation in Southeast Asia, 2000–2012. Proc. Natl. Acad. Sci. 113 (2), 344–349.
Roy, S.K., Deria, A., Hong, D., Rasti, B., Plaza, A., Chanussot, J., 2023. Multimodal fusion transformer for remote sensing image classification. IEEE Trans. Geosci. Remote Sens. 61, 1–20.
Rußwurm, M., Klemmer, K., Rolf, E., Zbinden, R., Tuia, D., 2023. Geographic location encoding with spherical harmonics and sinusoidal representation networks. arXiv preprint arXiv:2310.06743.
Rußwurm, M., Korner, M., 2017. Temporal vegetation modelling using long short-term memory networks for crop identification from medium-resolution multi-spectral satellite images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 11–19.
Schroeder, T.A., Schleeweis, K.G., Moisen, G.G., Toney, C., Cohen, W.B., Freeman, E.A., Yang, Z., Huang, C., 2017. Testing a Landsat-based approach for mapping disturbance causality in US forests. Remote Sens. Environ. 195, 230–243.
Shapiro, A., d'Annunzio, R., Desclée, B., Jungers, Q., Kondjo, H.K., Iyanga, J.M., Gangyo, F.I., Nana, T., Obame, C.V., Milandou, C., et al., 2023. Small scale agriculture continues to drive deforestation and degradation in fragmented forests in the Congo Basin (2015–2020). Land Use Policy 134, 106922.
Singh, C., Persson, U.M., 2024. Global patterns of commodity-driven deforestation and associated carbon emissions. http://dx.doi.org/10.31223/X5T69B, Published as preprint in the California Digital Library (CDL).
Sitzmann, V., Martel, J., Bergman, A., Lindell, D., Wetzstein, G., 2020. Implicit neural representations with periodic activation functions. Adv. Neural Inf. Process. Syst. 33, 7462–7473.
Slagter, B., Reiche, J., Marcos, D., Mullissa, A., Lossou, E., Peña-Claros, M., Herold, M., 2023. Monitoring direct drivers of small-scale tropical forest disturbance in near real-time with Sentinel-1 and-2 data. Remote Sens. Environ. 295, 113655.
Song, X.-P., Hansen, M.C., Potapov, P., Adusei, B., Pickering, J., Adami, M., Lima, A., Zalles, V., Stehman, S.V., Di Bella, C.M., et al., 2021. Massive soybean expansion in South America since 2000 and implications for conservation. Nat. Sustain. 4 (9), 784–792.
Tang, X., Bratley, K.H., Cho, K., Bullock, E.L., Olofsson, P., Woodcock, C.E., 2023. Near real-time monitoring of tropical forest disturbance by fusion of Landsat, Sentinel-2, and Sentinel-1 data. Remote Sens. Environ. 294, 113626.
Tang, F.H., Nguyen, T.H., Conchedda, G., Casse, L., Tubiello, F.N., Maggi, F., 2024. CROPGRIDS: a global geo-referenced dataset of 173 crops. Sci. Data 11 (1), 413.
Tropek, R., Sedláček, O., Beck, J., Keil, P., Musilová, Z., Šímová, I., Storch, D., 2014. Comment on "High-resolution global maps of 21st-century forest cover change". Science 344 (6187).
Tseng, G., Kerner, H., Rolnick, D., 2022. TIML: Task-informed meta-learning for agriculture. arXiv preprint arXiv:2202.02124.
Tyukavina, A., Hansen, M.C., Potapov, P., Parker, D., Okpa, C., Stehman, S.V., Kommareddy, I., Turubanova, S., 2018. Congo Basin forest loss dominated by increasing smallholder clearing. Sci. Adv. 4 (11), eaat2993.
Vancutsem, C., Achard, F., Pekel, J.-F., Vieilledent, G., Carboni, S., Simonetti, D., Gallego, J., Aragao, L.E., Nasi, R., 2021. Long-term (1990–2019) monitoring of forest cover changes in the humid tropics. Sci. Adv. 7 (10), eabe1603.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.
Wang, Y., Hollingsworth, P.M., Zhai, D., West, C.D., Green, J.M., Chen, H., Hurni, K., Su, Y., Warren-Thomas, E., Xu, J., et al., 2023. High-resolution maps show that rubber causes substantial deforestation. Nature 623 (7986), 340–346.
Zanaga, D., Van De Kerchove, R., Daems, D., De Keersmaecker, W., Brockmann, C., Kirches, G., Wevers, J., Cartus, O., Santoro, M., Fritz, S., Lesiv, M., Herold, M., Tsendbazar, N.-E., Xu, P., Ramoino, F., Arino, O., 2022. ESA WorldCover 10 m 2021 v200. Zenodo.