Multi-Spectral Segmentation of Deforestation Drivers in Sentinel-2 Imagery

Aurora Ingebrigtsen & Christian Bontveit & Synne Remmen Andreassen[1]

{ain015, zec019, xeb017}@uib.no

Abstract

This paper presents our approach to the Solafune competition "Identifying Deforestation Drivers", which involves classifying and segmenting deforestation drivers using Sentinel-2A satellite imagery. We detail our methodology for addressing the task, including the implementation of novel techniques to improve classification and segmentation performance. These include super-resolving the images, expanding the dataset with object based augmentation and fine-tuning a pretrained vision transformer. We evaluate our approach by comparing the results of our novel method pipelines against a baseline model, and we discuss the results and our achievements.

1 Introduction

Problem description

Deforestation — defined as the change of forested land to other types of land use, whether caused by human activity or natural processes — is a persistent and pressing environmental issue with wide-ranging ecological, climatic, and socio-economic consequences [1]. Accurately identifying and understanding the drivers of deforestation is a key objective in environmental monitoring and conservation.

This study is motivated by our participation in a machine learning competition organized by Solafune, a company focused on addressing social, environmental, and industrial challenges through geospatial technologies [2]. The competition's objective was to develop machine learning models capable of detecting and classifying the causes of deforestation based on satellite imagery.

Specifically, participants were tasked with building models that segment deforested areas into four driver categories: plantations, logging, mining, and grassland/shrubland. Each of these categories represents a distinct cause of land-use change, with unique spatial patterns and visual characteristics in satellite images.

We were provided with high-resolution multispectral data from the Sentinel-2 satellite. Sentinel-2 captures imagery across 12 spectral bands, covering the visible, near-infrared (NIR), and shortwave-infrared (SWIR) (portions of the electromagnetic) spectrum, providing detailed spectral information about the Earth's surface. An example Sentinel-2 image used in this study is shown below.

Related Work

Different versions of the deforestation problem have gained attention across several research fields in recent years. For example ForestNet was introduced to classify deforestation drivers in Indonesia, which a country has one of the highest deforestation rates in the world [3].

Satellite imagery analysis has become an active area of research within the machine learning and computer vision communities. A particular challenge is that different spectral bands in satellite imagery often have different ground sampling distances (GSD). In the case of Sentinel-2A which has bands that have GSD of 10m, 20m and 60m [4], the larger GSDs appear blurrier and contain less detail. Several approaches have been proposed to address this issue, ranging from statistical methods to machine learning-based techniques. One method is pan-sharpening [5], which uses a high-resolution panchromatic band to sharpen lower-resolution bands. Newer methods train neural networks to reconstruct high-resolution detail [6], however these methods may introduce hallucinations which may be undesirable in remote sensing [7]. This paper uses a novel method of super-resolution, bu using the high resolution band of an image to transfer texture [7].

Similarly, various preprocessing methods for satellite data have been proposed to enhance model performance. Data augmentation is a widely used technique to improve model generalization, especially in image-based tasks [8]. In the context of semantic segmentation, methods such as Cutout [9], MixUp [10], and CutMix [11] have been proposed to introduce additional variability during training by mixing or masking image regions. More recently, Copy-Paste [12] has been proposed, where objects are copied between images to generate new training samples. Building on these ideas, our work focuses on object-based augmentation made for remote sensing imagery.

Given the vast quantities of unlabeled satellite imagery, several self-supervised frameworks for transformer pre-training on multi-spectral imagery have been proposed. MAE [13] introduces a masked autoencoder architecture. This approach uses an encoder–decoder model that reconstructs randomly masked image patches to learn spectral–spatial features. SatMAE [14] adapts this approach to satellite data by employing spectral channel–aware masking. ConvMAE [15] injects convolutional inductive biases into both encoder and decoder. ScaleMAE [16] adds ground sample distance (GSD) based positional encodings to inform the model of the scale and location of the patches. Finally, SatMAE++ [17] extends SatMAE with multi-scale masking and reconstruction.

Objectives

The primary objective of this study is to develop a machine learning-based approach for land cover segmentation using high-resolution remote sensing data from Sentinel-2 imagery. Specifically, the goal is to accurately identify and localize the main drivers of deforestation across affected regions.

To this end, we implement and evaluate multiple convolutional neural network (CNN) architectures, including ResNet, assessing their performance based on validation metrics. Furthermore, we explore advanced data augmentation strategies and propose a novel model architecture aimed at enhancing segmentation accuracy and robustness.

Contributions

In this research, we developed segmentation models capable of classifying different deforestation drivers. We enhanced our models in several ways: by applying super-resolution techniques to improve image quality, by generating additional training samples through object-based augmentation, and by boosting model performance using a pretrained transformer architecture.

2 Methods

- What you did (choice of ML methods, use of ML methods, model selection/evaluation?)
- Reproducible

Exploratory Data Analysis

To begin the analysis, we first explored the dataset provided, which consisted of 176 satellite images along with their corresponding annotations. Each image contained 12 spectral bands, ordered as follows: Aerosols, Blue, Green, Red, Red Edge 1, Red Edge 2, Red Edge 3, Near-Infrared (NIR), Red Edge 4, Water Vapor, SWIR 1, and SWIR 2 (see 1 for an example image).

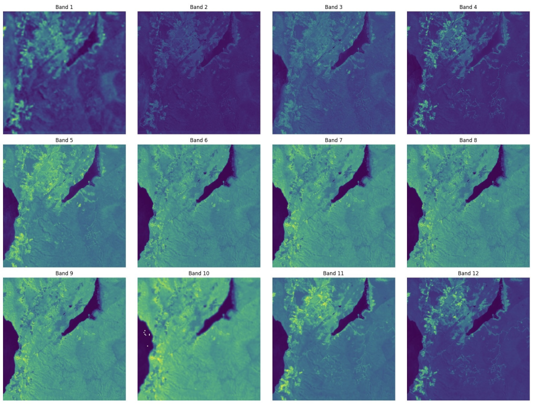

Figure 1. Plot of bands 1–12 for the image *train_1.tif*. The image contains examples of deforestation drivers, including logging, plantation, grassland/shrubland, and mining.

Some of the images contained NaN (Not a Number) values, primarily due to cloud coverage. Consequently, pre-processing steps were applied to handle these missing values, including their removal, followed by image normalization to standardize pixel intensities across all bands.

The dataset was subsequently divided into three subsets: 70% for training, 20% for evaluation, and 10% for testing. This resulted in a total of 123 images for the training set, which, while sufficient for initial experiments, presents a limitation in terms of model accuracy and generalization capability due to the relatively small sample size.

Evaluation and Selection

The evaluation metric for the task is IoU F1-score, as shown in figure 2. The IoU F1-score measures the balance between precision and recall by evaluating the overlap between the predicted segment and the ground truth.

$$F1 = 2 \cdot \frac{\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}}$$

Figure 2. The F1-score formula.

We designed a training pipeline where multiple network architectures a baseline convolutional network, a ResNet-based encoder-decoder and a novel transformer architecture (SatMAE) were defined and visualised. A grid search over key hyperparameters was performed for each architecture.

Models were trained on several augmented datasets, including super-resolved imagery and datasets enhanced through object-based augmentation techniques. After training, the best-performing model was selected based on validation F1-score. Finally, the selected model was evaluated on a separate, unseen test set to assess its generalization performance.

2.1 Super-resolution

The data we are working with consists of satellite imagery captured by the Sentinel-2 satellite. Each image contains 12 spectral bands, acquired at three different resolutions, referred to as ground sampling distances (GSDs) [4].

The available GSDs are 10m, 20m, and 60m, meaning that each pixel in the high-resolution bands represents a 10×10m area on the ground, while the lowest-resolution bands cover 60×60m per pixel. When visually comparing high- and low-resolution bands, it becomes evident that some detail is lost in the blur of the coarser resolutions.

The disparity in resolution across bands introduces a potential limitation: low-resolution bands may contain valuable information, but their coarser resolution can hinder accurate segmentation. Enhancing the lower-resolution bands to the same fidelity as the high-resolution ones, could improve the model's ability to accurately identify deforestation drivers. To address this, we explore a super-resolution approach that reconstructs high-resolution versions of the lower-resolution bands.

The super-resolution method implemented in this project was originally proposed by Lanaras et al. (2018) [7] and was at the time a novel approach to upsampling satellite images such as Sentinel-2. This method leverages the multi-band nature of Sentinel-2 images, by training networks to effectively transfer texture from the higher resolution bands to the lower resolution ones.

2.1.1 Implementation

The integration of the super-resolution method into our project is done as a pre-processing step, applied before the training pipeline begins. The super-resolution network is used to enhance the resolution of the lower-resolution bands prior to creating dataloaders. This ensures that all spectral bands are at the higher-resolution when passed to training.

While the original authors provided an implementation of their method in Keras/TensorFlow [18], our limited experience with this framework made it impractical for direct use. Fortunately, an unofficial PyTorch reimplementation was available online [19], which provided a suitable starting point.

However, this PyTorch implementation was designed to work specifically with the original Sentinel-2 .SAFE format, which is a hierarchical file structure containing not only the image bands but also extensive metadata [4]. Our data, provided by Solafune, was already processed into GeoTIFF raster format, which lacks this metadata structure. As a result, the available implementation could not be used out of the box.

We therefore had to develop our own adaptation of the method, compatible with the GeoTIFF format. This involved loading each GeoTIFF file and manually extracting and sorting the bands according to GSD and super-resolving them based on the higher-resolution bands.

Although the adaptation required significant modifications, the process was facilitated by the availability of pretrained model weights provided by Moskovchenko (2024) [19]. Without the pretrained model weights we would have had to train the super-resolution network from scratch, and for it to be adequate we would need a lot of data and compute. This enabled us to generate super-resolved outputs for use in the segmentation task.

2.2 Object-based augmentation

The performance of machine learning models for image segmentation tasks, such as identifying deforestation drivers in Sentinel-2 satellite imagery, is closely tied to the quality and quantity of training data. However, the availability of annotated satellite imagery suitable for this task is limited. Our dataset comprises 176 samples, which are partitioned into training, validation, and test subsets. The limited availability of data presents a challenge when training models with the goal of generalizing.

To mitigate this constraint, we adopted an object-based data augmentation approach, inspired by the method proposed by Illarionova et al. (2022) [20]. Unlike traditional augmentation techniques, object-based augmentation leverages segmentation masks to extract objects from the images. For our use case, such objects are areas indicative of deforestation activity. These objects are then reinserted into new spatial contexts within the image domain, generating realistic and diverse augmented samples.

The goal of this augmentation strategy is to enrich the training dataset and improve the model's generalization capabilities by introducing variations in spatial arrangements and environmental conditions associated with deforestation.

2.2.1 Implementation

As part of the data pre-processing pipeline, object-based augmentation was introduced as an additional step prior to dataset splitting. Before partitioning into training, validation, and test sets, each data sample was considered for augmentation with a predefined probability.

For selected samples, n objects were extracted from other labeled images in the dataset t, using their corresponding segmentation masks. These objects — representing relevant land features associated with deforestation — were then optionally augmented. With a given probability, geometric transformations (e.g., rotation, scaling) or color modifications were applied to the extracted objects. Although the augmentation framework allows for the artificial addition of shadows, this feature was not utilized, as our imagery involves natural landscapes rather than man-made structures. While the goal of including shadows is to enhance realism, applying shadows to land areas in our dataset decreased the realism of the samples. Therefore, shadows were excluded from our model but the functionality remains available for datasets where shadows contribute positively to realism.

Next, augmented objects were composited onto either a new background (sampled from another image) or the original background, selected probabilistically. Backgrounds themselves could also be augmented. Background and objects could be augmented separately or together. Objects were blended into the background using their original pixel values, and placement was constrained such that no object overlapped with another, and no object was duplicated within the same image.

2.3 Transformer Finetuning

As an option to the full model, a finetuned Vision Transformer (ViT-large) model was trained. The SatMAE++ framework [17] uses a MAE for pretraining, which reconstructs images at different scale levels. The multi scale reconstructions allows the model to learn large scale resolutions as well as fine grained details. We used the pretrained weights for the ViT-large model provided in `https://github.com/techmn/satmae_pp`. The weights are pretrained on the fMoW-full dataset, which is also sentinel-2 imagery. As our dataset was limited, no MAE pretraining on our dataset was done. For the finetuning a segmentation head was added to the ViT-large model, and a given number of layers was trained. The number of unfrozen layers was initially sat as a hyperparameter, but due to limited VRAM, training more than 2 of the layers was not computationally possible.

3 Results

Running the pipeline yielded the results in table 1.

Model	Normal	SR	OBA
Baseline (CNN)	0.5131	0.3026	0.5156
UNetResNet	0.3421	0.3947	0.4375
UNet	0.2236	0.1448	0.5357
ViT	0.1448	0.09	0.1875

Table 1. F1 score of each model ran for 30 epochs on each of the datasets using batch size 8, lr 0.001, decay 0.01 and momentum 0.9.

3.1 Super-resolution

The implementation of the super-resolution method was largely successful. As shown in Figure 3, the middle image represents the original 60m GSD band (B9), which appears significantly blurrier compared to the leftmost image — the original 10m GSD RGB composite (bands B2, B3, B4). On the rightmost side, we show the super-resolved version of the B9 band.

The super-resolved band is noticeably sharper, and when comparing it to the original 10m GSD RGB image, we observe that the details align well with the high-resolution. There are some noticable differences such as the handling of cloud-covered areas: in the original images, cloud regions are masked as white NaN values, whereas in our super-resolved outputs they appear black. The NaN-masked areas are substantially larger in the super-resolved output. While the exact cause is uncertain, one hypothesis is that the super-resolution model may have been confused by the presence of NaN regions during inference. Another observation is that the colors in the super-resolved images appear more subdued compared to the original, which could either be a result of the model behavior or related to our preprocessing before inference.

In conclusion, the super-resolution implementation successfully enhanced fidelity of the lower-resolution bands, achieving a sharper and more detailed output. However, some limitations, such as handling of NaN values and changes in color intensity.

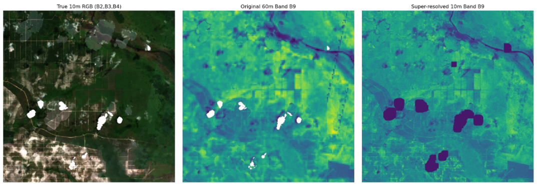

Figure 3. Plot of 6x super-resolution results *train_0.tif*. Leftmost 10m GSD rgb (B2, B3, B4), middle 20m GSD B9, rightmost super-resolved band B9.

Looking at the F1-scores in table 1 we can see how the models trained on super-resolved perfomed in the middle column. Using the baseline model it performed worst, using UNetResNet it perfomed second best, on UNet worst and on ViT it performed ...

3.2 Object-based augmentation

Object-based augmentation (OBA) led to higher F1 scores across all models. By generating additional training samples, OBA improved the models' ability to make more accurate and reliable predictions. As a result, models trained with OBA demonstrated better precision compared to those trained without it.

The models were trained using the default OBA parameters: three additional objects were added per image, no extra backgrounds were used, the probability of applying augmentation was set to 0.8, and the probability of performing object-based augmentation was set to 0.5. Although the highest F1 score achieved was 0.5357 when training a UNet model with OBA, the results suggest that object-based augmentation contributed to noticeable improvements in model performance

3.3 Transformer Pretraining

The pretrained Vision Transformer backbone used in our pipeline comprises 310,684,653 parameters in total. By freezing all layers except the final two transformer blocks, the subsequent layer normalization, and the classification head, the number of trainable parameters is reduced to 25,199,621. Attempting to unfreeze additional blocks was precluded by GPU memory constraints.

Figure 4 illustrates the training loss trajectory: after an initial drop, the loss rapidly stagnates, suggesting that limiting fine-tuning to only two transformer blocks may have constrained the model's capacity to adapt further to the downstream task.

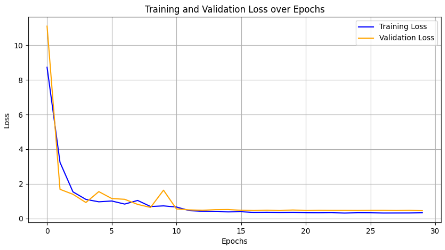

Figure 4. Vision Transformer loss over epochs

4 Discussion

Although the super-resolved bands appeared visually sharp and clear to the human eyes, this improvement in visual quality did not translate into better model performance. Models trained on the super-resolved datasets performed worse on some of the normal dataset trained models. Several factors may explain this outcome. First, the handling of NaN values during the super-resolution process may have introduced inconsistencies or artifacts in the data that negatively affected learning. Second, the super-resolved images exhibited more subdued colors compared to the original images, potentially altering the pixel value distributions and reducing the amount of useful spectral information available to the models. Together, these issues suggest that while super-resolution may improve human interpretability, but may mot improve machine learning results.

While it would have been interesting to experiment with additional background images for our object-based augmentation dataset, it is important to note that the environmental context cannot change drastically. One of the primary motivations for using new background images was to capture diverse environmental conditions, which may not have been feasible with our current dataset. However, had we obtained data from forests without significant deforestation, it is likely that we could have generated new, realistic training samples with the extracted objects for our model to improve its performance. This feature can be easily utilized within our implementation.

Shadows were not tested in the models due to time constraints and because they did not contribute to the realism of the images. The objective of adding shadows is to enhance realism, particularly for human-made objects. However, introducing shadows to land areas does not provide meaningful improvements. Although a method for adding shadows is provided in the implementation, it was not used for the models in this study.

Due to limited computing power, an extensive hyperparameter search was not conducted for all models, preventing us from thoroughly identifying the optimal model configuration. Additionally, we were unable to optimize the parameters for object-based augmentation and instead used the default settings provided in the original research. With more computational resources, we would explore additional models to gain a deeper understanding of the performance across our different pipelines.

- Interpretation
- Achieved objectives
- Compare with others

...

5 Conclusion

This paper presented our contributions to the Solafune deforestation drivers competition. We performed an exploratory data analysis on the deforestation sentinel-2 dataset provided. Further, we created a pipeline for training, validating and testing the models. We enhanced this pipeline by implementing three novel methods; super-resolving our images, increasing the size of our dataset by using object based augmentation, and by using a pretrained ViT using the SatMAE++ framework. Future work would include an extensive hyperparameter search. Also, further optimization of the dataset creation methods could be experimented with.

References

[1] Food and Agriculture Organization of the United Nations. *Global Forest Resources Assessment 2020 – Key Findings*. Accessed: 2025-04-21. Rome, 2020. URL: https://www.fao.org/3/ca8753en/CA8753EN.pdf.

[2] Solafune. *Solafune: A Platform for Remote Sensing AI Challenges*. https://solafune.com. Accessed: 2025-04-21. 2024.

[3] J. Irvin, H. Sheng, N. Ramachandran, S. Johnson-Yu, S. Zhou, K. Story, R. Rustowicz, C. Elsworth, K. Austin, and A. Y. Ng. "Forestnet: Classifying drivers of deforestation in indonesia using deep learning on satellite imagery". In: *arXiv preprint arXiv:2011.05479* (2020).

[4] C. SentiWiki. *Sentinel-2 Products*. https://sentiwiki.copernicus.eu/web/s2-products#S2Products-Level-2AProductsS2-Products-L2Atrue. Accessed: 2025-04-23. 2025.

[5] Q. Wang, W. Shi, and P. M. Atkinson. "Area-to-point regression kriging for pan-sharpening". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 114 (2016), pp. 151–165. ISSN: 0924-2716. DOI: https://doi.org/10.1016/j.isprsjprs.2016.02.006. URL: https://www.sciencedirect.com/science/article/pii/S0924271616000496.

[6] D. Pouliot, R. Latifovic, J. Pasher, and J. Duffe. "Landsat Super-Resolution Enhancement Using Convolution Neural Networks and Sentinel-2 for Training". In: *Remote Sensing* 10.3 (2018). ISSN: 2072-4292. URL: https://www.mdpi.com/2072-4292/10/3/394.

[7] C. Lanaras, J. Bioucas-Dias, S. Galliani, E. Baltsavias, and K. Schindler. "Super-resolution of Sentinel-2 images: Learning a globally applicable deep neural network". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 146 (2018), pp. 305–319. ISSN: 0924-2716. DOI: https://doi.org/10.1016/j.isprsjprs.2018.09.018. URL: https://www.sciencedirect.com/science/article/pii/S0924271618302636.

[8] Z. Wang, P. Wang, K. Liu, P. Wang, Y. Fu, C.-T. Lu, C. C. Aggarwal, J. Pei, and Y. Zhou. *A Comprehensive Survey on Data Augmentation*. 2024. arXiv: 2405.09591 [cs.LG]. URL: https://arxiv.org/abs/2405.09591.

[9] T. DeVries and G. W. Taylor. *Improved Regularization of Convolutional Neural Networks with Cutout*. 2017. arXiv: 1708.04552 [cs.CV]. URL: https://arxiv.org/abs/1708.04552.

[10] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. *mixup: Beyond Empirical Risk Minimization*. 2018. arXiv: 1710.09412 [cs.LG]. URL: https://arxiv.org/abs/1710.09412.

[11] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo. *CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features*. 2019. arXiv: 1905.04899 [cs.CV]. URL: https://arxiv.org/abs/1905.04899.

[12] G. Ghiasi, Y. Cui, A. Srinivas, R. Qian, T.-Y. Lin, E. D. Cubuk, Q. V. Le, and B. Zoph. *Simple Copy-Paste is a Strong Data Augmentation Method for Instance Segmentation*. 2021. arXiv: 2012.07177 [cs.CV]. URL: https://arxiv.org/abs/2012.07177.

[13] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. "Masked autoencoders are scalable vision learners". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 16000–16009.

[14] Y. Cong, S. Khanna, C. Meng, P. Liu, E. Rozi, Y. He, M. Burke, D. Lobell, and S. Ermon. "Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 197–211.

[15] P. Gao, T. Ma, H. Li, Z. Lin, J. Dai, and Y. Qiao. "Convmae: Masked convolution meets masked autoencoders". In: *arXiv preprint arXiv:2205.03892* (2022).

[16] C. J. Reed, R. Gupta, S. Li, S. Brockman, C. Funk, B. Clipp, K. Keutzer, S. Candido, M. Uyttendaele, and T. Darrell. "Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 4088–4099.

[17] M. Noman, M. Naseer, H. Cholakkal, R. M. Anwer, S. Khan, and F. S. Khan. *Rethinking transformers pre-training for multi-spectral satellite imagery*. 2024.

[18] C. Lanaras. *DSen2: Super-resolution for Sentinel-2 images*. https://github.com/lanha/DSen2. Accessed: 2025-04-21. 2018.

[19] M. Moskovchenko. *simonreise/remote-sensing-processor: Remote Sensing Processor 0.2.2 (v0.2.2)*. 2024. DOI: 10.5281/zenodo.11091321. URL: https://doi.org/10.5281/zenodo.11091321.

[20] S. Illarionova, S. Nesteruk, D. Shadrin, V. Ignatiev, M. Pukalchik, and I. Oseledets. *Object-Based Augmentation Improves Quality of Remote Sensing Semantic Segmentation*. 2022. arXiv: 2105.05516 [cs.CV]. URL: https://arxiv.org/abs/2105.05516.