A TRANSFORMER-BASED NETWORK FOR DEFORESTATION DETECTION FROM BITEMPORAL OPTICAL SATELLITE IMAGES

Felipe Ferrari[1], *Matheus Pinheiro Ferreira*[2] *Raul Queiroz Feitosa*[3]

[1]Pontifical Catholic University of Rio de Janeiro (PUC-Rio), ferrari@aluno.puc-rio.br,
[2]Military Institute of Engineering (IME), matheus@ime.eb.br,
[3]Pontifical Catholic University of Rio de Janeiro (PUC-Rio), raul@ele.puc-rio.br

ABSTRACT

The expansion of deforestation rates in the Brazilian Amazon has raised attention to forest monitoring initiatives. Currently, the PRODES project employs human specialists to detect new deforestation areas from bitemporal optical satellite images. Previous works showed that convolutional neural networks (CNNs) achieved excellent results to automatically detect deforestation. Recently, models based on Swin Transformers layers outperformed the CNN models in many computer vision tasks. In this paper, we investigated the employment of the Swin Transformer-based model to detect new deforestation areas, compared to a traditional CNN-based network. Swin Transformer-based models increased the F1-Score by 0.055*. The code can be found here.*

Key words *– Deep Learning, Remote Sensing, Deforestation Detection, Transformer.*

1. INTRODUCTION

As the largest tropical rainforest on Earth, the Amazon Forest provides essential ecosystem services such as biodiversity harboring and climate regulation [1]. However, deforestation and forest have impacted these ecosystem services [2]. In 2021, for example, Brazil lost 13,235 km^2 of forests in the Amazon [3].

Since the 1980s, the National Institute for Space Research (INPE) has monitored deforestation in the Brazilian biomes through the PRODES (Project of Deforestation Monitoring by Satellite) project, relying on optical images. However, the deforestation assessment by PRODES still involves much visual interpretation, with undesirable implications in terms of time and costs. Semi-automatic approaches have been explored to minimize such drawbacks with minimum accuracy loss [3].

Deep Learning emerged in the past few years as a promising approach to this goal. In particular, fully convolutional networks (FCNs) have achieved state-of-the-art results in several remote sensing tasks [4], including deforestation detection [5]. A typical FCN architecture consists of an encoder module for feature extraction, followed by a decoder module that delivers a prediction at the input image resolution. Many FCN variants have been applied for deforestation detection from optical images [6–8] as well as combined optical and SAR data [9].

About two years ago, an alternative to FCNs, called Vision Transformers or simply Transformers, arose as an alternative to FCNs for image dense prediction tasks [10]. A significant limitation of FCNs derives from its fundamental operation, the convolution, which restricts the calculation of a new representation of each pixel to a limited spatial context. In contrast, Vision Transformers forego convolutions and can explore the input's global context. Indeed, the last two years have witnessed an increasing number of Transformer architectures achieving superior performance than FCNs in dense prediction tasks such as semantic segmentation [11].

However, as far as we know, no work has been published to date to evaluate Transformers for automatic deforestation detection in Brazilian biomes within a PRODES-like context.

This work aims to fill this gap and presents the first comparison between a state-of-the-art FCN architecture with a Transformer based model for deforestation detection in selected sites of the Brazilian Legal Amazon. Specifically, we compare the Swin-Unet Transformer, initially conceived for semantic segmentation of medical images [12], with an FCN based on convolutional Residual Blocks representing the current SOTA on the deforestation detection task [9]. Both models are Encoder-Decoder architectures with skip connections, as detailed in the following.

The remaining of this work is organized as follows. Section 2 presents the material and methods, explaining the models employed. Section 3 shows the study area, described the dataset, and the experimental protocol. Section 4 summarizes the findings of the current study.

2. MODELS' ARCHITECTURES DESCRIPTION

2.1. Swin Transformer-based model

Figure 1a presents the overall Swin-Unet architecture. Initially, the Input Image $X \in \mathbb{R}^{H\times W\times B}$ is partitioned into image tokens $\in \mathbb{R}^{4\times 4\times B}$, where H, W, and B are the height, the width, and the number of bands in the Input Image, respectively. A Linear Embedding layer projects each feature map onto an arbitrary C-depth.

The Swin-Unet Encoder comprises a sequence of Swin Transformer Blocks followed by a Patch Merging operation. Each Swin Transformer Block consists of a Windowed Multi-head Self-Attention (W-MSA) or the Shifted W-MSA (SW-MSA) operation. The W-MSA and SW-MSA apply the MSA between image tokens restricted to the same window. To ensure the connection between the windows, each W-MSA is followed by an SW-MSA, which shifts the window by a fixed number of tokens. Following the W-MSA or SW-MSA, there are two layers of Multilayer Perceptron (MLP), with Gaussian Error Linear Unit (GELU) activation modules. Each module is preceded by a Layer Normalization (LN) operation and a residual connection, as shown in Figure 1b.

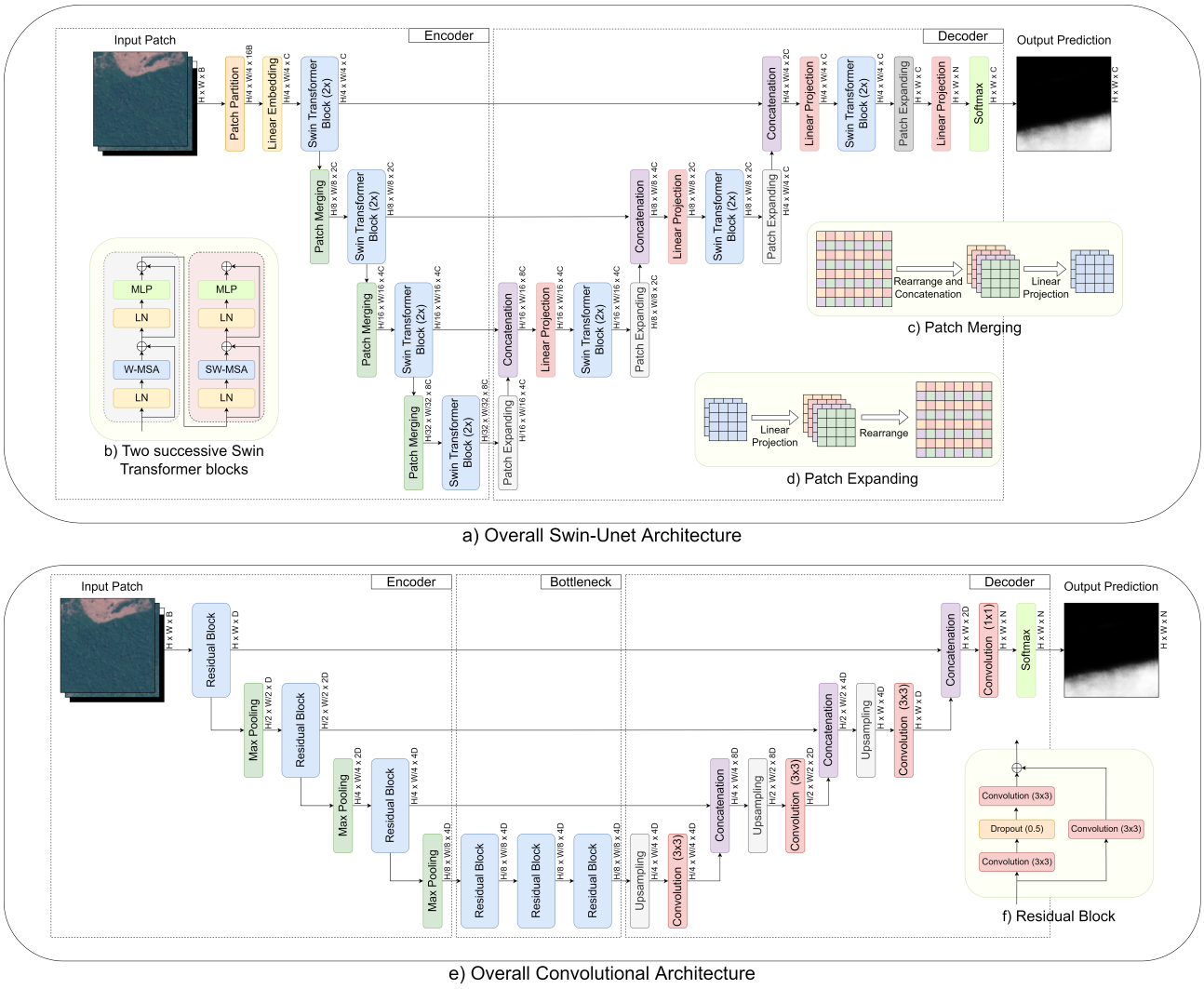

Figure 1: Models' Architectures: a) Overall Swin-Unet architecture, b) Two successive Swin Transformer blocks, c) Patch Merging, d) Patch Expanding, e) Overall CNN-based architecture, and f) Residual Block.

Patch Merging is applied to reduce the height and width of the image tokens while expanding the depth. Each Patch Merging operation halves the height and width while doubling the depth dimension. This is done by a linear projection of the concatenated rearranged image tokens, as presented in Figure 1c.

The Swin-Unet Decoder consists of a Patch Expanding operation, a concatenation with the respective Encoder output, a Linear Projection, to reduce the depth, and Swin Transformer Blocks, applied in sequence until the tokens' height and width turn back to their original size. The last Patch Expanding procedure is then applied, to recover the original image height and width from the image tokens, followed by a linear projection and softmax activation to generate the Output Prediction with N classes, as shown in Figure 1a.

Patch Expanding performs the opposite operations of Patch Merging, increasing image tokens' depth with a linear projection followed by an image tokens rearrangement, increasing the height and width while reducing the depth, as presented in Figure 1d. All Patch Expanding doubles the size while halves the depth dimension, except the last one, which quadruples the size without changing the depth dimensionality.

2.2. CNN-based model

The CNN-based model consists of an Encoder, a Bottleneck, and a Decoder, as shown in Figure 1e. The Encoder consists of a sequence of Residual Blocks followed by 2×2 max pooling operations. The first Residual Block's convolutions have D filters and the following doubles after each max pooling operation. The Bottleneck is a sequence of three Residual Blocks producing feature maps with the same depth $4D$. The Decoder comprises an upsampling operation, followed by a 3×3 convolutional layer, whose outcome is concatenated with the feature map produced by the corresponding Encoder until the height and width match the original image size. Finally, a 1×1 convolutional layer, with a softmax activation, produces the Output Prediction with N classes.

The Residual Block is a sequence of two 3×3 convolutional layers with a dropout between them and a residual connection, in which a 3×3 convolutional layer is applied just to adjust the depth dimensionality. Figure 1f shows the Residual Block. After 3×3 convolutional layers, a ReLU activation is applied.

3. EXPERIMENTAL ANALYSIS

3.1. Reference data

We used the deforestation areas identified by PRODES as reference data [3]. PRODES employs trained photo-interpreters to manually outline deforested areas on satellite images. Only areas larger than 6.25 hectares where the primary forest was completely removed are outlined. For visual interpretation, the specialists select images with a minimum cloud cover and acquired within the dry season (July to September) [3].

We derived the pixel-wise labels for all the tiles from PRODES deforestation polygons available at the TerraBrasilis platform [13]. We labeled the image pixels as *No Deforestation* if no deforestation has been detected until t_0, *Deforestation* if deforestation was detected between t_0 and t_1, and *Background* for areas deforested before t_0 and no forest features.

Due to differences in the spatial resolution between the images used to generate the PRODES polygons, Landsat images with 30 m spatial resolution, and this work, Sentinel with 10 m spatial resolution, a buffer with 2 pixels was applied to *Deforestation* polygons.

3.2. Satellite images and study area

In this work, we used bitemporal (t_0-t_1) Sentinel-2 images, with 13 Level-1C bands from each image, concatenated to the previous deforestation temporal distance map (explained later), generating a 27 bands image, henceforth called Input Image. All bands were resampled to 10 meters, when necessary, and normalized to mean 0 and standard deviation 1. We trained the models on images acquired in 2017 and 2018 and tested them on images from 2018 and 2019. In all cases, we selected Sentinel-2 images acquired close to the Landsat counterparts used for the corresponding PRODES report.

The previous deforestation is a piece of information that we already know from PRODES and can help the identification of new future deforestation areas. This information can be presented as a temporal distance map, in which the pixels can assume values between 0 and 1. If no previous deforestation was identified in a pixel, this value is 0. If a previous deforestation was identified in a pixel, its value depends on the year of the identification. If the identification year is t_0, its value is 1, and it linearly decreases as far from t_0 is the identification year.

The area of this work is in the State of Pará, Brazil. We split the study area Input Image (2017-2018) into training and validation tiles to train the models. We used the entire study area Input Image (2018-2019) to evaluate the models' predictions. This study area has an area of $\approx 16,300 Km^2$.

3.3. Experimental protocol

All procedures of this work were conducted with an Intel Core i7-7800X processor, with 64 GB of RAM and a GeForce GTX 1080 Ti GPU with 11 GB of dedicated memory and Python language.

The Swin-Unet and CNN-based architectures have the C and D, respectively, arbitrary values, which affects the number of trainable parameters. Our study set $C = 96$ and $D = 32$, following [12] and [9] respectively.

We generated patches of size 128×128 pixels from the training and validation tiles with 70% overlap from the Input Image (2017-2018). To reduce the imbalance inherent to the deforestation detection task, we split the patches into two groups, depending on whether it has more or equal to 2% of *Deforestation* pixels or less than that. For the training and validation sets, we selected randomly the same number of patches from each of those groups.

To minimize the effect of class imbalance, we used a weighted binary cross entropy (WBCE) loss function with weights of 0.1, 0.9, and 0 for the *No Deforestation*, *Deforestation*, and *Background* classes, respectively. We chose the Adam optimizer with a learning rate of $1 \cdot 10^{-4}$. We trained 10 models from scratch for each proposed architecture until there was no improvement in the loss function for ten consecutive epochs in the validation data.

Finally, we performed predictions on the testing data, Input Image (2018-2019), using patches of size 128×128 pixels with an overlap of 16 pixels between them. Each prediction patch was cropped, keeping only the central region with 96×96 pixels. From these cropped patches, the prediction of the whole image was generated. The deforestation predictions of each trained model were evaluated by F1-Score, precision, and recall, with a 0.5 threshold value.

3.4. Results and discussion

The proposed CNN-based architecture has nearly 2.72 M of parameters, while Swin-Unet has approximately 27.19 M of parameters ($\approx 10\times$), as shown in Figure 2c. However, the larger number of parameters from Swin-Unet did not significantly increase the training and prediction times, as presented in Figures 2a and 2b. The Swin-Unet models' mean training and prediction times were 80% and 53% greater than CNN-based models.

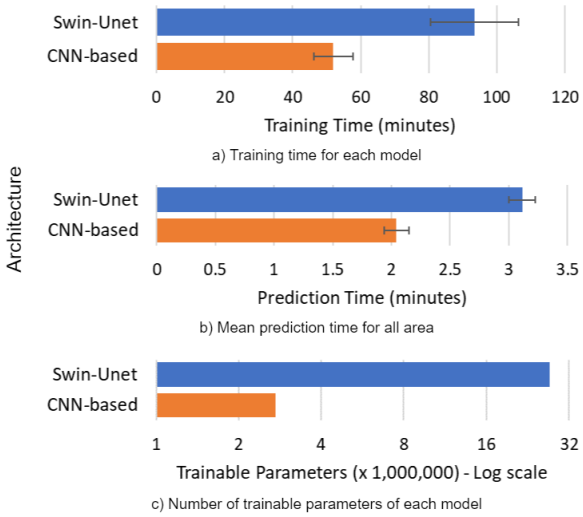

Figure 2: Processing times and the number of trainable parameters.

The mean and the standard deviation of the evaluated metrics of each deforestation prediction for each model

architecture can be found in Figure 3. The Swin-Unet showed greater results than CNN-based for F1-Score and precision metrics but presented a slightly lower recall mean value.

The standard deviation of the CNN-based models' predictions presented greater values than the Swin-Unet for all evaluated metrics.

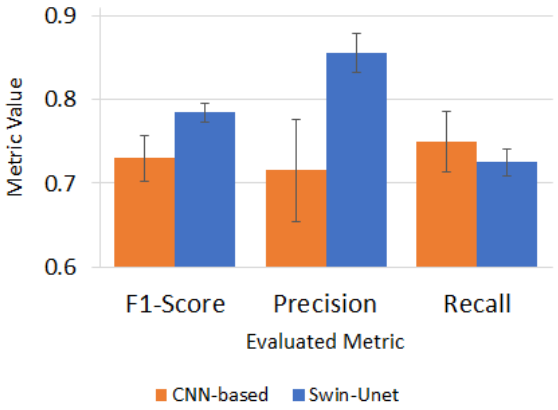

Figure 3: Mean and standard deviation of the metrics.

Figure 4 shows the prediction samples of a region of the study area from CNN-based (a) and Swin-Unet (b) models, respectively, and the respective reference data (c). Both models can identify larger deforestation areas. However, the CNN-based model generates more false positives, highlighted by the yellow arrows in Figure 4, confirming the difference in the precision metric between the models.

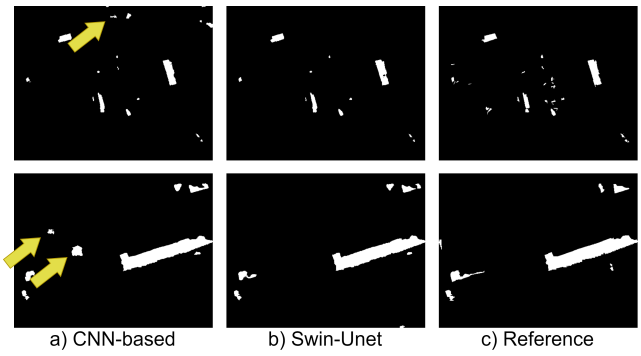

Figure 4: Prediction samples and the respective reference.

4. CONCLUSIONS

The current study aimed to investigate if Swin Transformer-based models, rising as state-of-the-art architecture, can achieve the same quality for identifying new deforestation areas tasks.

Extrapolating the experiments' times to the Brazilian Amazon Forest area, the estimated times increased from 11 days to 20 days, and from 10 hours to 16 hours, for training and prediction, respectively. Despite the increase, it does not make its utilization unfeasible, especially if powerful hardware is available.

From all the presented results, the Swin-Unet models produced better results than CNN-based considering all metrics except the recall. More investigations are necessary, especially about Swin Transformer-based architectures, which still demand many trainable parameters. However, these models showed promising results to aid the deforestation monitoring task.

5. REFERENCES

[1] J. Strand, B. Soares-Filho, M. H. Costa, U. Oliveira, S. C. Ribeiro, G. F. Pires, A. Oliveira, R. Rajao, P. May, R. van der Hoff, *et al.*, "Spatially explicit valuation of the brazilian amazon forest's ecosystem services," *Nature Sustainability*, vol. 1, no. 11, pp. 657–664, 2018.

[2] A. Baccini, W. Walker, L. Carvalho, M. Farina, D. Sulla-Menashe, and R. Houghton, "Tropical forests are a net carbon source based on aboveground measurements of gain and loss," *Science*, vol. 358, no. 6360, pp. 230–234, 2017.

[3] INPE, "Metodologia utilizada nos sistemas PRODES e DETER," 2022.

[4] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, "Deep learning in remote sensing applications: A meta-analysis and review," *ISPRS journal of photogrammetry and remote sensing*, vol. 152, pp. 166–177, 2019.

[5] M. O. Adarme, R. Q. Feitosa, P. N. Happ, C. A. D. Almeida, and A. R. Gomes, "Evaluation of deep learning techniques for deforestation detection in the brazilian amazon and cerrado biomes from remote sensing imagery," *Remote Sensing*, vol. 12, p. 910, 3 2020.

[6] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual u-net," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, pp. 749–753, 5 2018.

[7] D. L. Torres, J. N. Turnes, P. J. S. Vega, R. Q. Feitosa, D. E. Silva, J. M. Junior, and C. Almeida, "Deforestation detection with fully convolutional networks in the amazon forest from landsat-8 and sentinel-2 images," *Remote Sensing*, vol. 13, p. 5084, 12 2021.

[8] P. P. de Bem, O. A. de Carvalho, R. F. Guimarães, and R. A. T. Gomes, "Change detection of deforestation in the brazilian amazon using landsat data and convolutional neural networks," *Remote Sensing*, vol. 12, 3 2020.

[9] M. X. Ortega, R. Q. Feitosa, J. D. Bermudez, P. N. Happ, and C. A. D. Almeida, "Comparison of optical and sar data for deforestation mapping in the amazon rainforest with fully convolutional networks," in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pp. 3769–3772, IEEE, 7 2021.

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 10 2020.

[11] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," 3 2021.

[12] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," 5 2021.

[13] L. F. F. G. Assis, K. R. Ferreira, L. Vinhas, L. Maurano, C. Almeida, A. Carvalho, J. Rodrigues, A. Maciel, and C. Camargo, "Terrabrasilis: A spatial data analytics infrastructure for large-scale thematic mapping," *ISPRS International Journal of Geo-Information*, vol. 8, p. 513, 11 2019.