# Deep Transformer-Based Network Deforestation Detection in the Brazilian Amazon Using Sentinel-2 Imagery

Mariam Alshehri, Anes Ouadou, and Grant J. Scott, *Senior Member, IEEE*

***Abstract*— Deforestation poses a critical environmental challenge with far-reaching impacts on climate change, biodiversity, and local communities. As such, detecting and monitoring deforestation are crucial, and recent advancements in deep learning (DL) and remote sensing technologies offer a promising solution to this challenge. In this study, we adapt ChangeFormer, a transformer-based framework, to detect deforestation in the Brazilian Amazon, employing the attention mechanism to analyze spatial and temporal patterns in bitemporal satellite images. To assess the model's effectiveness, we employed a robust approach to create a deforestation detection (DD) dataset, utilizing Sentinel-2 imagery from select conservation areas in the Brazilian Amazon throughout 2020 and 2021. Our dataset comprises 7734 pairs of bitemporal image chips with a resolution of 256 × 256 pixels and 1406 pairs of image chips with a resolution of 512 × 512 pixels. The model achieved an overall accuracy (OA) of 93% with a corresponding F1 score of 90% and an intersection over union (IoU) score of 82%. These results demonstrate the potential of transformer-based networks for accurate and efficient DD.**

***Index Terms*— Change detection (CD), deep learning (DL), deforestation, transformer.**

## I. Introduction

DEFORESTATION significantly impacts environmental sustainability, causing biodiversity loss, ecological imbalances, and amplified climate change effects. The Brazilian Amazon, the world's largest rainforest, is indispensable for climate stability and carbon management. Unfortunately, rapid deforestation leads to multiple complications, including heightened greenhouse gas emissions, curtailed carbon retention, and increased forest fires [1]. Therefore, it is essential to implement effective policies that are grounded in reliable, up-to-date data, recognizing deforestation detection (DD) as the cornerstone for obtaining such valuable information.

Although DD is recognized as an essential task in restoring the biodiversity of the Brazilian Amazon, it is fraught with various challenges. One of the major obstacles is the vast size of the Brazilian Amazon, covering approximately 5.2 million km$^2$ of land, which is about 60% of the country's total area [2]. This makes conventional methods, such as map interpretation, field surveys, and ancillary data analysis, impractical due to their time-consuming and labor-intensive nature.

Remote sensing imagery (RSI) has become a more advanced alternative, owing to its wide geographic coverage, cost-effectiveness, and capacity to produce consistent and reproducible data, vital for tracking temporal changes. The popularity RSI has increased further with the introduction of open-access policies for Earth observation satellites and advances in analytical technologies [3], leading to the creation of multiple RSI systems by both public and private organizations.

Moderate Resolution Imaging Spectroradiometer (MODIS), Landsat 8, and Sentinel-2 are some of the most commonly used satellites in remote sensing studies. MODIS offers 36 spectral bands with a maximum spatial resolution of 250 m and a revisit time of two days. Landsat 8 has 11 spectral bands with 15-m panchromatic and 30-m multispectral spatial resolutions and a revisit time of 16 days. Sentinel-2, launched by European Space Agency (ESA), has 13 spectral bands with a spatial resolution range of 10–60 m and a revisit time of five days. Sentinel-2's advantage lies in its higher spatial and temporal resolution compared with the other open-access satellites, making it a better option for mapping the expansive Brazilian Amazon.

In RSI change detection (CD) applications, a quantitative analysis is performed to identify surface changes by comparing images of the same location captured at different timestamps. The aim is to accurately detect pixel changes in bitemporal or multitemporal images by assigning a binary label to each pixel. A null label represents an unchanged area, while a positive label indicates the presence of change. Various CD methods have been proposed in the literature, including algebra-based techniques (such as image ratioing and image difference) and machine learning classifiers (such as support vector machines, decision trees, and fuzzy theory) [4], [5]. However, these methods fail to produce accurate results when high-resolution images are used due to the high contrast and frequency components of the images.

To address the limitations of the existing methods, deep learning (DL) methods have emerged as dominant techniques for image analysis. The literature indicates that DL methods demonstrate superior performance compared with traditional machine learning techniques in CD applications [6], [7]. Convolutional neural networks (CNNs) are a widely used

Manuscript received 10 May 2023; revised 1 November 2023; accepted 10 January 2024. Date of publication 17 January 2024; date of current version 21 May 2024. This work was supported in part by Princess Nourah Bint Abdulrahman University, Riyadh, Saudi Arabia, and in part by the National Science Foundation (NSF) under Grant OAC-1925681. *(Corresponding author: Mariam Alshehri.)*

Mariam Alshehri is with the Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO 65211 USA, and also with the College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University, Riyadh 84428, Saudi Arabia (e-mail: msapwz@umsystem.edu).

Anes Ouadou and Grant J. Scott is with the Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO 65211 USA.

Digital Object Identifier 10.1109/LGRS.2024.3355104

TABLE I
TOP-5 CONSERVATION UNITS RANKED BY DEFORESTATION SIZE

| Area ID and name | Area(km$^2$) | Deforestation(km$^2$) | Deforestation(%) |
|---|---|---|---|
| 59-Área de Proteção ambiental Triunfo do Xingu | 16,792 | 528.33 | 3.14 |
| 89-Floresta Nacional do Jamanxim | 13,017 | 214.59 | 1.63 |
| 55-Reserva Extrativista Jaci-Paraná | 1,974 | 108.92 | 5.51 |
| 291-Área de Proteção Ambiental do Tapajós | 20,395 | 114.70 | 0.56 |
| 165-Reserva Extrativista Chico Mendes | 9,315 | 91.31 | 0.98 |

classification method in CD applications, due to their ability to generate powerful discriminative features. For example, [1] utilized CNNs to predict annual changes in vegetation cover within the Brazilian Amazon; however, this technique entails redundant operations, leading to increased computational costs. The performance of CNNs for CD has been enhanced by integrating dilated convolutions, stacked convolution layers, and attention mechanisms into the models.

The transformer model, a DL model incorporating attention mechanisms, is well suited for accommodating multitemporal images, as it enables easy scaling, captures long-range sequence features, and supports efficient parallel processing. Considering these benefits, the transformer model has been applied in various areas of computer vision, such as the vision transformer (ViT) [8], bitemporal image transformer (BIT) [9], and shifted windows (SWin) transformer [10]. One of the key advantages of these networks over CNNs is that they offer superior context-modeling ability between pixel pairs, because they have a larger effective receptive field [11]. Regardless of the potential of transformer models, there is limited research regarding their application in DD. This study aims to investigate the application of the transformer-based network, ChangeFormer [12], originally designed for CD in urban areas, in the context of deforestation, with the objective of demonstrating its effectiveness in DD. The model combines a hierarchical transformer encoder in a Siamese architecture, four difference modules for computing feature differences, and a simple MLP decoder. The anticipation is that this transformer-based network will provide superior DD accuracy compared with that of CNNs.

## II. METHODS

### A. Data Sources

*1) Satellite Imagery Source:* Sentinel-2 satellite images were used due to their superior spatial and temporal resolution compared with other open-access satellites. Images were downloaded from The Copernicus Open Access Hub[1] operated by ESA. We used Level-2A Surface Reflectance product with a maximum cloud cover percentage of 20%. The images were downloaded in tiles, each of size 10 980 × 10 980 pixels, covering an area of approximately 100 × 100 km.

*2) Ground-Truth Source:* To generate ground-truth polygons for our study, we utilized the PRODES project datasets developed by the Brazilian National Space Agency (INPE). Since 1988, the PRODES project has been monitoring and quantifying annual deforestation rates in the Brazilian Amazon rainforest by visually interpreting medium-resolution satellite imagery with a team of experienced professionals. The

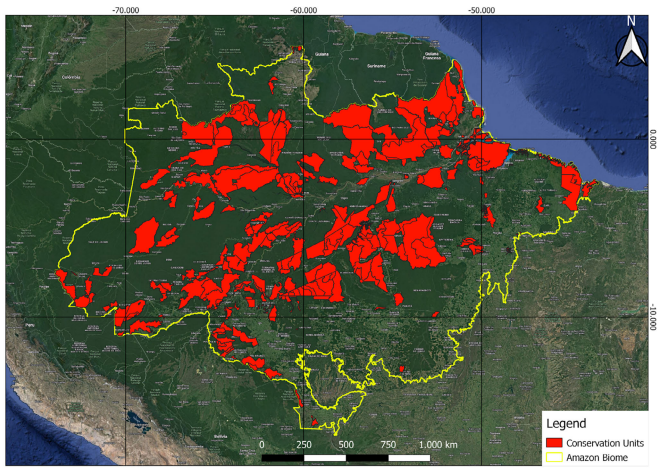


Fig. 1. Amazon biome is outlined by the yellow boundary, and the conservation units are highlighted in red.

PRODES data are publicly accessible on the TerraBrasilis website.[2] We selected two datasets: the yearly deforestation dataset, which contains the locations where deforestation occurred from one year to the next, and the conservation units dataset, which we used to identify the areas from where we downloaded the images. Fig. 1 shows the boundaries of Amazon biome and the conservation units.

### B. Location and Date Selection

The top-5 conservation units with the highest deforested land areas were selected. Table I details their names, respective sizes in km$^2$, and the corresponding deforestation areas and percentages. We targeted the years 2020 and 2021, as they provide the most recent and complete yearly deforestation data. To determine the appropriate date range for each Sentinel-2 tile, we identified the polygons within the tile for both years 2020 and 2021 from the PRODES dataset. For each year separately, we extracted the earliest date ($d_1$) and the latest date ($d_2$) from the acquisition dates of the images used to label these polygons. Consequently, two intervals were defined for each tile: one for 2020 and another for 2021, each set to $[d_1 - 30$ days, $d_2 + 30$ days$]$. These extended intervals aim to capture a variety of images for quality selection while ensuring that the imagery used in our study aligns closely with the ground-truth labeling dates, minimizing the likelihood of significant deforestation developments occurring in the interim period between image acquisition and labeling.

### C. Band Combinations

Sentinel-2 images are composed of 13 bands at varying resolutions of 10, 20, and 60 m. We used the visible

[1]Copernicus portal: https://scihub.copernicus.eu/dhus/

[2]Terrabrasilis download: http://terrabrasilis.dpi.inpe.br/en/download-2/

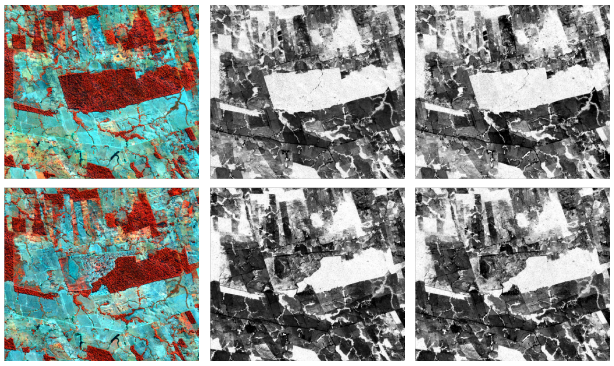

Fig. 2. Temporal and spectral comparison of satellite images from (top) 2020 and (bottom) 2021 highlighting deforestation changes. The first column shows NGB images with forest appearing in red, while nonforest areas appear in blue-green tones. The second and third columns show the same area but with NDVI and EVI indices, respectively.

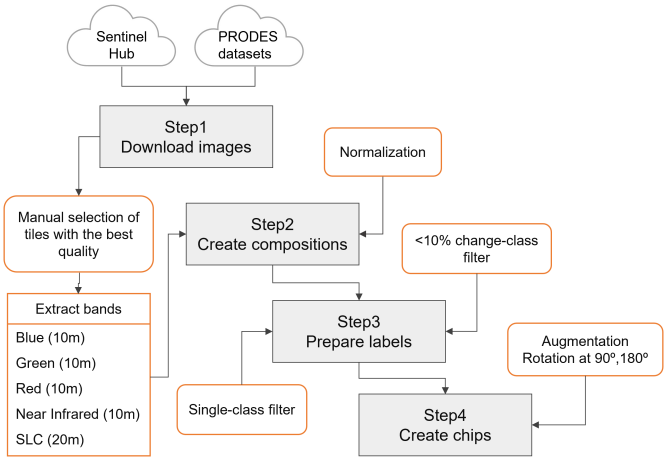


Fig. 3. Dataset creation process.

and near-infrared bands at 10-m resolution, along with the scene classification layer (SCL) at 20-m resolution. The near infrared is effective in distinguishing vegetation from other features, as healthy vegetation reflects strongly in the near-infrared part of the spectrum. The SCL band was resampled from 20 to 10 m to match the resolution of the other selected bands and used as a mask to remove areas classified as clouds or cloud shadows, which were labeled as "no change." We used the following band combinations that emphasize the spectral signature of vegetation.

1) *Color-Shifted Infrared:*

$$\text{NGB} = [\text{Band8}, \text{Band3}, \text{Band2}].$$

2) *Normalized Difference Vegetation Index:*

$$\text{NDVI} = \frac{\text{Band8} - \text{Band4}}{\text{Band8} + \text{Band4}}.$$

3) *Enhanced Vegetation Index:*

$$\text{EVI} = 2.5 * \frac{\text{Band8} - \text{Band4}}{\text{Band8} + 6\ \text{Band4} - 7.5\ \text{Band2} + 1}.$$

Fig. 2 illustrates the spectral variations resulting from different band combinations applied to a pair of images captured in 2020 and 2021.

### D. Image Processing

To ensure that our dataset is representative of the problem domain and free from potential biases, we employed careful preprocessing procedures. Our first step involved applying linear normalization to facilitate image comparison on a consistent scale. The minimum and maximum values were selected as the 1st and 99th percentiles of the value histogram to reduce the impact of outliers. This step ensures that the resulting pixel values span the full range of possible values for that band, which can help to enhance the contrast and quality of the image. We also applied filtering at both the raster and chip levels. High-quality rasters were manually selected from the returned results of the download query. At the chip level, we eliminated single-class chips and those where the "change" class accounted for less than 10% of the chip's total area to mitigate extreme class imbalance between change and no-change classes. Since deforestation changes in one year are often dispersed across large no-change areas, the 10% filter resulted in a significant reduction of chips. To compensate for this loss, we implemented rotation augmentation at 90° and 180° to increase the dataset size and enhance model robustness and generalization, and further augmentation as described in [13] could be implemented as needed.

At the end of the chip creation process, a total of 7734 pairs of chips of size 256 × 256 and 1406 pairs of chips of size 512 × 512 were generated. The dataset then was split into three subsets: 60% for training, 20% for testing, and 20% for validation. To ensure the integrity of the dataset, we took measures to ensure that each chip and its rotations were kept within the same subset, and the distribution of the classes is maintained in each subset. The dataset creation process is summarized in Fig. 3, highlighting the key steps.

## III. Experiments

In the employed ChangeFormer architecture, a hierarchical transformer encoder processes bitemporal images, generating ConvNet-like multilevel features through self-attention modules and downsampling blocks. The key component of this architecture comprises four difference modules. These modules calculate feature differences between prechange and postchange images at multiple levels, employing the following sequence of operations: $F^i_{\text{diff}} = \text{BN}(\text{ReLU}(\text{Conv2D}^{3\times3}(\text{Cat}(F^i_{\text{pre}}, F^i_{\text{post}}))))$, where $F^i_{\text{diff}}$ represents the feature difference, BN stands for batch normalization, ReLU is the rectified linear unit function, $\text{Conv2D}^{3\times3}$ denotes a 2-D convolution with a 3 × 3 kernel, and Cat indicates the concatenation of $F^i_{\text{pre}}$ and $F^i_{\text{post}}$, which are the features of prechange and postchange images, respectively. These feature differences are then aggregated by a simple MLP decoder to predict the change map. The decoder encompasses MLP layers and upsampling steps to fuse feature difference maps, producing the final change mask prediction. Refer to Fig. 4 for a simplified diagram and the original research paper for comprehensive details [12].

We randomly initialized the model and optimized the performance by tuning six hyperparameters listed in Table II. We trained multiple model configurations from the

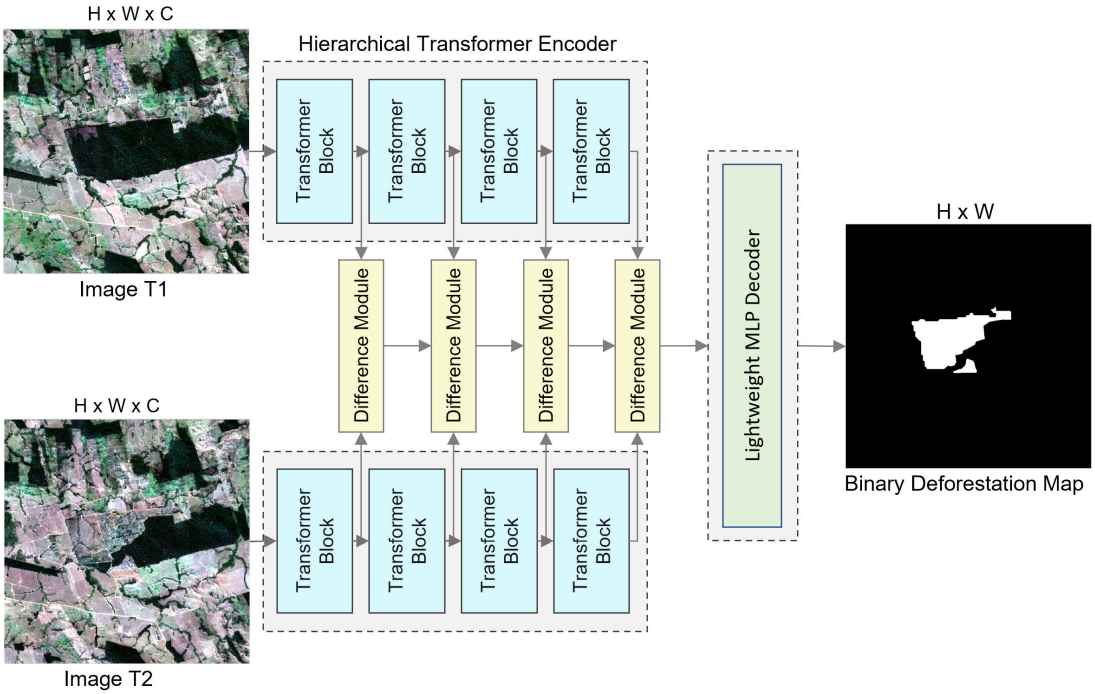


Fig. 4. Simplified overview of the ChangeFormer architecture, showing its three main components: a Siamese hierarchical transformer encoder, four difference modules, and a lightweight MLP decoder.

TABLE II
HYPERPARAMETERS INVOLVED IN TUNING PROCESS

| Hyperparameter | Variations |
|---|---|
| Image size | 256, 512 |
| Band composite | NGB, EVI, NDVI |
| Optimizer | SGD, Adam, AdamW |
| learning rates | 0.01 - 0.0001 |
| Loss functions | Cross-Entropy (CE), Intersection over Union (IoU) |
| Batch size | 4, 8, 12, 16 |

hyperparameter space for 200 epochs each and evaluated their performance using both overall and change-class-specific metrics. Overall metrics, including mean F1 score, mean intersection over union (IoU) score, and overall accuracy (OA), were reported alongside change-class-specific metrics, which included F1, IoU, precision, and recall. We utilized Nautilus, a high-performance computing system, to perform our data processing pipeline and model training with multiple configurations.

## IV. RESULTS

Although OA is a widely used metric, it may not be the best indicator for our deforestation dataset due to the inherent class imbalance. Therefore, to provide a more comprehensive evaluation, we considered mIoU as an overall metric for comparison. In addition, we reported change-class-specific metrics, including F1, IoU, precision (pre-1), and recall (rec-1) scores, to further assess the model's ability to detect deforestation changes. This focus on change-class metrics obtains accurate representation of the model's ability to detect deforestation changes, rather than results that might be skewed by the detection of the majority unchanged class.

The 256 × 256 chips exhibited better performance in DD, effectively capturing smaller deforested patches and their surrounding context, thereby generating a more diverse pool of training samples from the same geographic area and enhancing model generalization. Conversely, the broader spatial context of the 512 × 512 chips often includes a larger proportion of the "no-change" class. While these chips can capture more deforested patches, the extensive area they cover often results in these patches constituting less than 10% of the chip area, triggering their elimination based on the data filter steps. This disproportionately reduces the sample size for the 512 × 512 chips, indirectly elevating the risk of overfitting due to narrowed variability in the data.

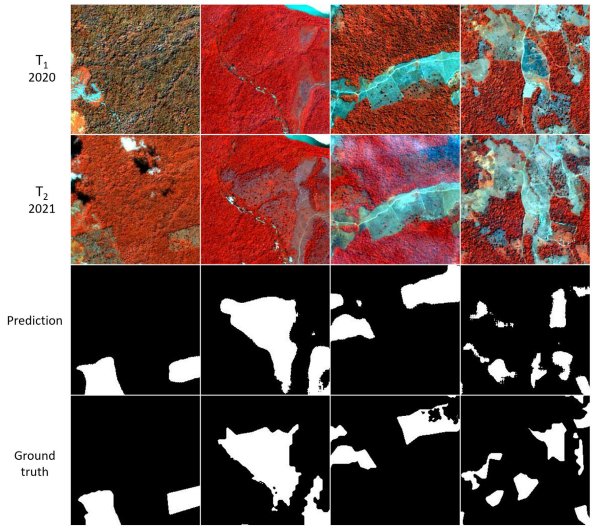


Fig. 5. Four image samples with the corresponding predicted and ground-truth deforestation maps. Top-2 rows show images from 2020 and 2021, respectively. The third and fourth rows display the predicted and ground-truth maps, respectively.

TABLE III
TOP-5 SCORING HYPERPARAMETER CONFIGURATIONS FOR (TOP) 256 × 256 AND (BOTTOM) 512 × 512 CHIP SIZES

| chip size | optimizer | loss | lr | OA | mIoU | mF1 | IoU-1 | F1-1 | pre-1 | rec-1 |
|---|---|---|---|---|---|---|---|---|---|---|
| | AdamW | CE | 0.0001 | 0.9306 | **0.8238** | **0.9007** | **0.7333** | **0.8462** | **0.8470** | **0.8453** |
| | Adam | CE | 0.0001 | 0.9286 | 0.8196 | 0.8980 | 0.7274 | 0.8422 | 0.8407 | 0.8437 |
| 256 | AdamW | CE | 0.001 | 0.9243 | 0.8077 | 0.8902 | 0.7080 | 0.8290 | 0.8459 | 0.8128 |
| | Adam | mIoU | 0.0001 | 0.9227 | 0.8072 | 0.8901 | 0.7098 | 0.8303 | 0.8229 | 0.8378 |
| | AdamW | mIoU | 0.0001 | 0.9221 | 0.8067 | 0.8898 | 0.7095 | 0.8301 | 0.8174 | 0.8432 |
| | AdamW | CE | 0.0001 | **0.9336** | 0.8008 | 0.8842 | 0.6788 | 0.8087 | 0.8262 | 0.7918 |
| | Adam | CE | 0.0001 | 0.9322 | 0.7980 | 0.8824 | 0.6750 | 0.8059 | 0.8179 | 0.7944 |
| 512 | AdamW | mIoU | 0.0001 | 0.9306 | 0.7956 | 0.8808 | 0.6721 | 0.8039 | 0.8051 | 0.8026 |
| | Adam | mIoU | 0.0001 | 0.9273 | 0.7904 | 0.8775 | 0.6657 | 0.7993 | 0.7826 | 0.8168 |
| | AdamW | CE | 0.001 | 0.9223 | 0.7759 | 0.8674 | 0.6421 | 0.7821 | 0.7771 | 0.7871 |

We found that the combination of a learning rate of 0.0001, CE loss, AdamW optimizer, and the NGB band produced the best results within our experimentation. Table III presents the results categorized by chip size (256 × 256 and 512 × 512), displaying the top-5 scores for each image size. Fig. 5 shows sample bi-temporal images with the predictions of the top-performing model.

A previous study [14] applied DL for DD by comparing six state-of-the-art fully convolutional network architectures, namely, U-Net, ResU-Net, SegNet, FC-DenseNet, Xception, and MobileNetV2 variants of Deeplabv3+. Similar to our work, they used the PRODES dataset for ground truth, and they utilized both Sentinel 2 and Landsat-8 satellite imagery. The best result of this prior study for Sentinel-2 data was FC-DenseNet with an F1 score of 70.7%. However, the ChangeFormer model obtained at least 80.9% F1 score for the change class (F1-1) across both chip sizes. Similar trends were observed in recall (rec-1) and precision (pre-1), with the values of 84.5% and 84.7%, respectively, outperforming the 75.1% and 78.0% reported in the previous study.

## V. CONCLUSION AND FUTURE WORK

We obtained Sentinel-2 satellite imagery and ground-truth data for deforested areas in the Brazilian Amazon rainforest in 2020 and 2021. Using these data, we created a bitemporal deforestation dataset and trained a transformer-based network for DD. We conducted a thorough hyperparameter search, exploring various configurations to identify the best settings for our task. Our investigation showed that color-shifted infrared composite and cross-entropy loss with AdamW optimizer resulted in the highest mean IoU (0.9007) and mean F1 score (0.8238) with precision and recall of the change class of 84.70% and 84.53%, respectively, demonstrating a well-balanced detection performance. In comparison with the existing research in the field of DD, our findings suggest that transformer-based networks are capable of achieving significantly improved levels of accuracy.

Our future work involves experimenting with state-of-the-art methods on a larger dataset that we plan to make publicly available along with the data acquisition and processing pipeline code. This could be a valuable contribution to the field of DD, enhancing the accuracy and robustness of DL models in this domain.

## ACKNOWLEDGMENT

Mariam Alshehri would like to acknowledge the support provided by her primary affiliation, Princess Nourah Bint Abdulrahman University, Riyadh, Saudi Arabia.

## REFERENCES

[1] P. de Bem, O. de Carvalho Junior, R. F. Guimaraes, and R. T. Gomes, "Change detection of deforestation in the Brazilian Amazon using Landsat data and convolutional neural networks," *Remote Sens.*, vol. 12, no. 6, p. 901, Mar. 2020.

[2] C. A. Silva, G. Guerrisi, F. Del Frate, and E. E. Sano, "Near-real time deforestation detection in the Brazilian Amazon with Sentinel-1 and neural networks," *Eur. J. Remote Sens.*, vol. 55, no. 1, pp. 129–149, Dec. 2022.

[3] M. Lu, E. Pebesma, A. Sanchez, and J. Verbesselt, "Spatio-temporal change detection from multidimensional arrays: Detecting deforestation from MODIS time series," *ISPRS J. Photogramm. Remote Sens.*, vol. 117, pp. 227–236, Jul. 2016.

[4] M. Ortega Adarme, R. Queiroz Feitosa, P. Nigri Happ, C. Aparecido De Almeida, and A. Rodrigues Gomes, "Evaluation of deep learning techniques for deforestation detection in the Brazilian Amazon and cerrado biomes from remote sensing imagery," *Remote Sens.*, vol. 12, no. 6, p. 910, Mar. 2020.

[5] F. Pan, Z. Wu, X. Jia, Q. Liu, Y. Xu, and Z. Wei, "A temporal-reliable method for change detection in high-resolution bi-temporal remote sensing images," *Remote Sens.*, vol. 14, no. 13, p. 3100, Jun. 2022.

[6] S. H. Khan, X. He, F. Porikli, and M. Bennamoun, "Forest change detection in incomplete satellite images with deep neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 9, pp. 5407–5423, Sep. 2017.

[7] Y. Liu, X. Chen, Z. Wang, Z. J. Wang, R. K. Ward, and X. Wang, "Deep learning for pixel-level image fusion: Recent advances and future prospects," *Inf. Fusion*, vol. 42, pp. 158–173, Jul. 2018.

[8] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," Jun. 2021, *arXiv:2010.11929*.

[9] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1-14, 2022, Art no. 5607514

[10] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. ICCV*, Oct. 2021, pp. 10012–10022.

[11] K. Han et al., "A survey on vision transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 87–110, Jan. 2023.

[12] W. G. C. Bandara and V. M. Patel, "A transformer-based Siamese network for change detection," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2022, pp. 207–210.

[13] G. J. Scott, M. R. England, W. A. Starms, R. A. Marcum, and C. H. Davis, "Training deep convolutional neural networks for land-cover classification of high-resolution imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 4, pp. 549–553, Apr. 2017.

[14] D. L. Torres et al., "Deforestation detection with fully convolutional networks in the Amazon forest from Landsat-8 and Sentinel-2 images," *Remote Sens.*, vol. 13, no. 24, p. 5084, Dec. 2021.