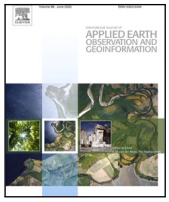
International Journal of Applied Earth Observation and Geoinformation 120 (2023) 103332

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

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



TransU-Net++: Rethinking attention gated TransU-Net for deforestation mapping☆

Ali Jamali [a,1], Swalpa Kumar Roy [b,1], Jonathan Li [c], Pedram Ghamisi [d,e,*]

[a] Department of Geography, Simon Fraser University, British Columbia 8888, Canada
[b] Department of Computer Science and Engineering, Jalpaiguri Government Engineering College, West Bengal 735102, India
[c] Department of Geography and Environmental Management, University of Waterloo, Ontario, ON N2L 3G1, Canada
[d] Helmholtz-Zentrum Dresden-Rossendorf (HZDR), Helmholtz Institute Freiberg for Resource Technology, 09599 Freiberg, Germany
[e] Institute of Advanced Research in Artificial Intelligence (IARAI), 1030 Vienna, Austria

ARTICLE INFO

Keywords:
Deforestation mapping
Vision transformer
Deep learning
Segmentation
Forest degradation
U-Net

ABSTRACT

Deforestation has become a major cause of climate change, and as a result, both characterizing the drivers and estimating segmentation maps of deforestation have piqued the interest of researchers. In the computer vision domain, Vision Transformers (ViTs) have shown their superiority compared to extensively utilized convolutional neural networks (CNNs) over the last couple of years. Although, ViTs has several challenges, specifically in remote sensing image processing, including their significant complexity that increases the computation costs and their need for much higher reference data than that of CNNs. As such, in this paper, we introduce an attention gates aided TransU-Net, called **TransU-Net**++ for semantic segmentation with an application of deforestation mapping in two South American forest biomes, i.e., the Atlantic Forest and the Amazon Rainforest. The heterogeneous kernel convolution (HetConv), U-Net, attention gates, and ViTs are all utilized in the proposed **TransU-Net**++ to their advantage. The **TransU-Net**++ significantly increased the performance of TransU-Net's over the Atlantic Forest dataset by about 4%, 6%, and 16%, respectively, in terms of overall accuracy, F1-score, and recall, respectively.Moreover, the results show that the developed TrasnU-Net++ model (0.921) achieves the highest Area under the ROC Curve value in the 3-band Amazon forest dataset as compared to other segmentation models, including ICNet (0.667), ENet (0.69), SegNet (0.788), U-Net (0.871), Attention U-Net-2 (0.886), R2U-Net (0.888), TransU-Net (0.889), Swin U-Net (0.893), ResU-Net (0.896), U-Net+++ (0.9), and Attention U-Net (0.908), respectively. The code will be made publicly available at https://github.com/aj1365/TransUNetplus2.

1. Introduction

Forests, which cover a wide range of landforms, constitute tree branches, herbs, vegetation, and numerous species of animals, including mammals, algae, bacteria, and other life forms. Forests approximately cover 30% of the surface of the Earth and has a substantial effect in the global environment and climatic conditions. Forest cover is universally recognized to be crucial to the conservation of biodiversity, carbon capture, watershed protection, climate change mitigation (Bonan, 2008), bioclimatic equilibrium, precipitation level maintenance (Pires and Costa, 2013), and the long-term viability of broad climatological regions (Boers et al., 2017). In brief, forests provide a diverse set of ecosystem functions and livelihood opportunities for human-being (Lausch et al., 2016; Schulze et al., 2019), making them beneficial for their financial, environmental, and recreation activities roles, as well as playing a significant impact on Earth's atmosphere patterns (Etteieb et al., 2013; Vanhala et al., 2005). For example, the Amazon rainforest accounts for roughly 5.6 million km^2, accounting for 50% of the remaining tropical forestland on the planet, and it affects significantly to global and local climate stabilization (Alzu'bi and Alsmadi, 2022; Maslin et al., 2005). Moreover, it is suggested that the forest of the Amazon is a crucial part of a large-scale moisture system that can be permanently impacted by forest loss (Boers et al., 2017). Although forests are considered as an essential component of the ecological system, they are degraded for a range of factors. Population growth worldwide and the rising urbanization, together with the degradation due to economic actions, have resulted in a significant shift

☆ This research was funded by the Institute of Advanced Research in Artificial Intelligence (IARAI).
* Corresponding author at: Helmholtz-Zentrum Dresden-Rossendorf (HZDR), Helmholtz Institute Freiberg for Resource Technology, 09599 Freiberg, Germany.
E-mail addresses: alij@sfu.ca (A. Jamali), swalpa@cse.jgec.ac.in (S.K. Roy), junli@uwaterloo.ca (J. Li), p.ghamisi@gmail.com (P. Ghamisi).
[1] Ali Jamali and Swalpa Kumar Roy contributed equally to this work.

https://doi.org/10.1016/j.jag.2023.103332
Received 31 December 2022; Received in revised form 7 April 2023; Accepted 25 April 2023

(i.e., forest loss) in global forested land, particularly in tropical zones such as Amazon forests of Brazil (Malhi et al., 2008).

Moreover, deforestation globally has risen dramatically in recent years due to various factors, including natural catastrophes and anthropogenic activities, such as energy generation and construction. For instance, in accordance with the National Institute for Space Research (NISR), the loss of forest of Amazon increased drastically in the 1990s (Müller, 2020), prompting numerous concerns about the disappearance of the Amazon rainforest by 2030. The Amazon Rainforest accounts for roughly 40% of the world's surviving tropical forests (Hubbell et al., 2008). As a result, the Amazon Rainforest's massive carbon sequestering functionality is critical to the regulation of the global and continental environment, as it is predicted to capture carbon by about 76 billion tonnes from 390 billion trees (Müller, 2020). The Atlantic Forest biome is Brazilian's economic driver, accounting for 70% of the country's GDP, 2/3 of its industrialized economy, and home to several of the greatest productive farmlands (Scarano and Ceotto). The Atlantic Forest became one of the "hottest" ecological regions as a result of the historic loss of habitat and breakdown brought on by urban development, industrial growth, and agricultural production (Laurance, 2009; Rezende et al., 2018).

Nevertheless, South America's explosive growth has already experienced significant forest loss due to agricultural production and land for constructions (García-Ayllón, 2016). As a result, the influence of deforestation contributes to global warming (Mikhaylov et al., 2020). Forest loss accounts for nearly 10% of emissions of Earth's greenhouse gas (GHG), and average Earth's temperatures have risen by over 1 degree Celsius as of pre-industrial periods (Samset et al., 2020). As such, delivering high-quality, precise information in a timely manner regarding changes in forest cover at various scales is crucial to improved surveillance and monitoring of forest degradation. This entails the creation and implementation of techniques designed to deal with the massiveness of forest cover zones as well as the high recurrence and level of detail of existing satellite imagery in a reasonable time frame, which also signifies that the manual operations and maintenance task must be lowered to the maximum extent possible. Field investigation and photo interpretation approaches have historically been used to observe tropical forests (Bragagnolo et al., 2021a). The primary constraint of such techniques is their labor-intensive characteristics that also necessitate exhaustive human influence (Gong et al., 1994). Consequently, the enormity of Brazil's tropical forests renders them unaffordable. Nevertheless, technological improvements in remote sensing have largely driven the creation of a variety of methodologies used in forest regions surveillance, including Decision Trees (Hansen et al., 2013), Random Forest (Ahmad et al., 2022; Yin et al., 2017), Regression Trees (Sexton et al., 2013), and Maximum Likelihood (Hamunyela et al., 2017). The main limitation of the traditional methods, such as the Random Forest, is their manual feature extraction. As a result, the success of traditional algorithms significantly relies on the proper selection of features (i.e., feature engineering). This issue was addressed by the deep learning techniques with their automatic feature extraction.

Deep learning has lately piqued the interest of remote sensing scientists because of its capability to obtain discriminative characteristics from images, such as satellite images (Jamali et al., 2022b; Rasti et al., 2020; Ghamisi et al., 2019; Jamali et al., 2022b,a). It is being used in a variety of complex surroundings and tasks, yielding significant outcomes in image classification (Roy et al., 2020; Ma et al., 2019; Roy et al., 2020; Roy et al., 2021) and object detection (Zhao et al., 2019; Sandhya Devi et al., 2021). Nevertheless, one drawback of deep learning models such as convolutional neural networks (CNNs) is that it only indicates the likelihood of a feature of desire appearing in an image but does not deliver knowledge of where the feature appears in the image. As such, Long et al. recommended a fully convolutional network (FCN) that permits the labeling of each pixel as a specific class, i.e., a pixel-based image classifying method, named semantic segmentation, as compared to image labeling, to overcome this restriction and enlarge the variety of uses of CNNs. The semantic segmentation process is additionally one of the difficult yet efficient techniques used in image understanding and analysis (Hao et al., 2020), medical image processing (Hesamian et al., 2019), and remote sensing (Kemker et al., 2018; Qi et al., 2020).

The U-Net algorithm, introduced by Ronneberger et al. (2015) that initially was employed in biomedical imaging, is one of the many FCN structures. To improve segmentation performance, this structure concatenates feature maps at various levels. The primary difference of the U-Net model and traditional FCNs is explained as the greater quantity of expansion networks, which permits the model to propagate data to layers of higher resolution. The U-Net structure has been used effectively in various studies, including remote sensing (Wang et al., 2022a; Zhang et al., 2021a; John and Zhang, 2022). Moreover, an attention-based U-Net algorithm was tested and discussed for forest loss in Amazon that showed improvement of the U-Net model via utilization of attention mechanism (John and Zhang, 2022). Waldeland et al. (2022) employed a U-Net algorithm for forest monitoring taking advantage of multispectral imagery of Sentinel-2 in Africa.

On the other hand, the CNNs' intrinsic backbone restrictions prevent them from accurately capturing the sequential characteristics of satellite images' spectral reflectance. This issue can be addressed by the proper utilization of the self-attention mechanism with vision transformers (ViTs) algorithms. besides, the inclusion of ViTs with CNN networks has been shown to be a powerful tool for downstream computer vision tasks (Wang et al., 2022c; Dutta et al., 2022; Gulzar and Khan, 2022; Lu et al., 2023; Jamali et al., 2023). For instance, for semantic image segmentation, Wang et al. employed and concluded that the integration of ViTs with the CNN architecture will considerably improve the segmentation results as compared to solo utilization of CNNs (Wang et al., 2022c). The use of ViTs as the backbone for CNN-based architectures for image segmentation was proposed and discussed by Dutta et al. that illustrated the significant improvement of the results with the inclusion of the ViTs (Dutta et al., 2022). Moreover, medical image segmentation research by Gulzar and Khan demonstrated the superiority of the integration of CNNs with ViTs as compared to the use of only CNN models (Gulzar and Khan, 2022). Wang et al. explored a lightweight ResNet-18 encoder to extract an efficient global–local attention mechanism to model both global and local information in the decoder by constructing an UNet-like transformer (UNetFormer) for real-time urban scene segmentation (Wang et al., 2022b). Yuan et al. introduced a CNN and Transformer as a complementary network for medical image segmentation it improves feature representation ability as well as achieves superior performance on multi-organ and cardiac image segmentation tasks. (Yuan et al., 2023).

As such, we propose the TransU-Net++, an attention gates aided TransU-Net (Chen et al., 2021), which incorporates the capabilities of heterogeneous kernel convolution (HetConv), U-Net, Attention gates, and ViTs for semantic segmentation. We propose a deep learning U-Net-based semantic segmentation structure and made beneficial use of the HetConv operation to capitalize heterogeneous kernels inside the learning units for degradation-free feature representation learning. Moreover, the utilization of attention gates substantially improves the positional information extraction of the decoder network. In addition, the use of HetConv operations and attention gates significantly improves the segmentation performance as compared to its baseline TransU-Net algorithm.

This paper introduces a semantic segmentation framework the TransU-Net++ for deforestation in the Atlantic and Amazon forests region in Section 2, illustrates the experiments and result analysis in Section 3, discusses the spatial transferability of the segmentation models in Section 4, and highlights the concluding remarks in Section 5.

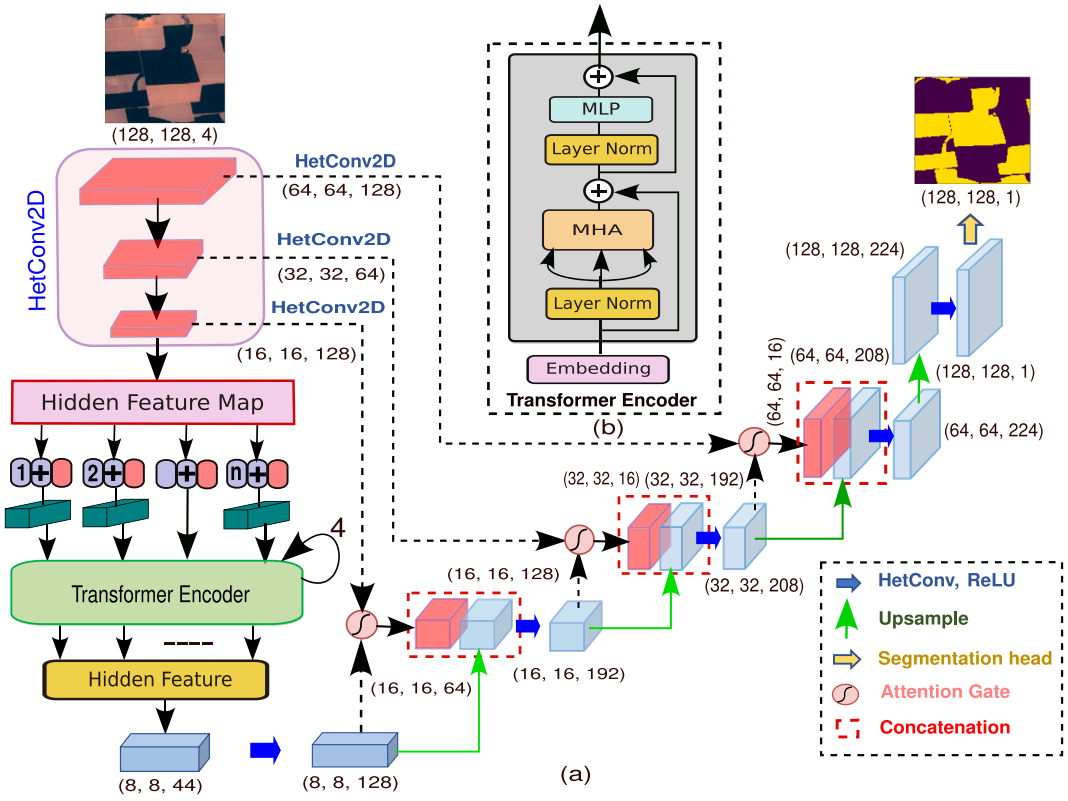


Fig. 1. Graphical representation of the proposed architecture (a) TransU-Net++ and (b) Transformer encoder.

2. Materials and methods

2.1. Datasets and experimental settings

We utilized three sets of satellite imagery from SentinelHub to assess the capability and efficiency of the TransU-Net++ algorithm for deforestation mapping. The first set of data is a 3-band dataset, which includes RGB-converted satellite images and forest loss masks of the Amazon Rainforest (Bragagnolo et al., 2021b,a). The other two datasets are 4-band data containing RGB and near-infrared (NIR) imagery of the Atlantic and Amazon rainforests (Bragagnolo et al., 2021b,a), as seen in Fig. 2. We utilized 400 training images in the two datasets of the Amazon and Atlantic forests with four bands. It should be noted that images have a $512 \times 512 \times C$ shape, where C denotes the number of color bands. Each forest loss mask has a shape of $512 \times 512 \times 1$. The dataset producer divided a significant number of satellite images into sub-images and generated masks by employing a customized k-means model with the use of the software of GRASS-GIS 7.6.1 (Bragagnolo et al., 2021b,a). In the first dataset, there are 30 images as training (we used 21 images as training and 9 images as validation datasets) and 15 images as validation data, where we used the validation images as our test data. In the other two 4-band datasets, there are 499, 100, and 20 images as training, validation, and test data, respectively. All experiments were done in an Intel Core-i7 CPU and NVIDIA RTX 2070 MAX-Q GPU in Python programming language. It should be noted that all segmentation models have been developed in Tensorflow 2.7 with a learning rate, batch size, and epoch of 0.001, 1, and 40, respectively, and the training is performed using the Adam optimizer. For the implemented models, the input images had a dimension of $128 \times 128 \times C$. It should also be noted that we utilized the mentioned image as input size even though it will increase the computation costs of some of the evaluated segmentation models, including Swin U-Net.

2.2. Proposed methodology

Considering a remote sensing image $\mathbf{X} \in R^{H \times W \times C}$ where H and W define spatial height and width and C presents the number of color channels, respectively. The objective is to define $\mathbf{Y} = \mathcal{F}(\mathbf{X})$ that represents the pixel label output map (i.e., classification map) of input $\mathbf{X}$ of the same spatial size of $(H \times W)$. The basic process is to utilize a convolutional U-network, which maps input images into high-level feature representations in the encoding steps and then decodes it back to a full spatial resolution to generate a pixel-wise label classification map. As shown in Fig. 1, we introduce the TransU-Net++ as an improved version of the original TransU-Net model for deforestation segmentation, which incorporates elements of HetConv, U-Network, attention gates, and vision transformers (ViT). There are several major benefits to using the TransU-Net++: (1) heterogeneous kernel convolution (HetConv) incorporates point-wise convolution and group-wise convolution to optimize the effectiveness of the generalized representation; (2) skip connections that link a low-level feature to its corresponding high-level feature, which improve knowledge transmission without deteriorating it, allowing us to create a lower complexity structure that obtains increased semantic segmentation information with a minimum amount of reference data; (3) ViT is a model based on self-attention and has excellent ability to capture long-range features dependencies, as well as obtaining global correlation; and (4) attention gates that improve the positional information and learn to emphasize on target objects of various sizes and shapes.

2.3. Description of the used models

Fig. 1 shows the overall of the proposed TransU-Net++ network, which is detailed in the following sections. The individual components of the TransU-Net++ network are detailed in the following subsections:

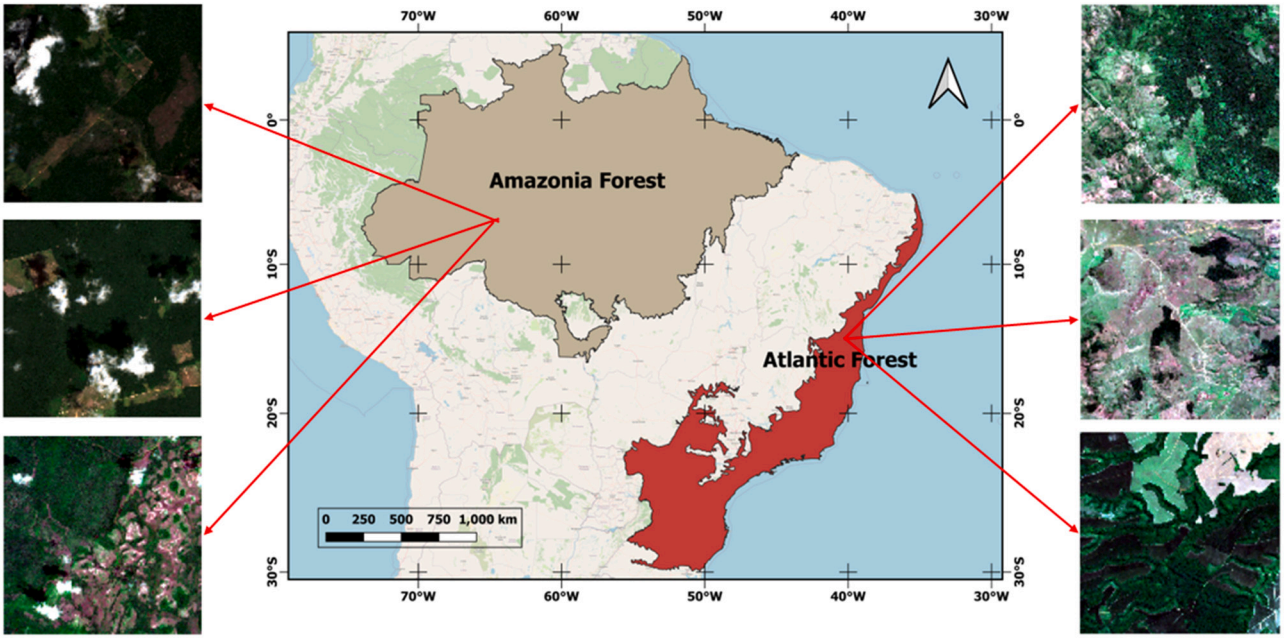


Fig. 2. Map of the 4-bands Atlantic Forest and 3- and 4-bands Amazon Rainforest biomes located in Brazil, South America.

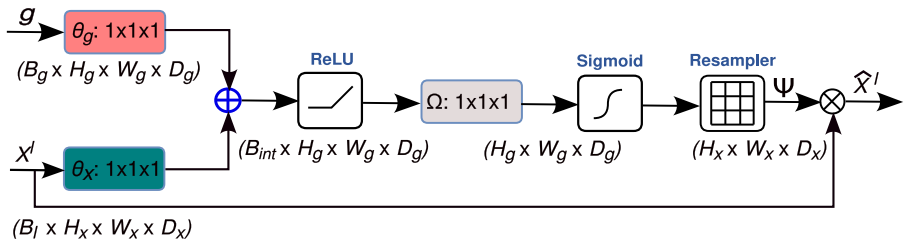


Fig. 3. Schematic representation of attention gates (AG).

2.3.1. Transformer encoder

The image input of $\mathbf{X}$ is reshaped into a set of flattened 2D patches, $x_p^i \in R^{P^2 \times C}\{i = 1, 2, \ldots, N\}$, where each patch has a size of $(P \times P)$ and the number of image patches are defined as N. The vectorized patches x_p are mapped into a latent embedding space of d-dimensional by employing a linear projection layer that can be trained. The position embedding is also added with the spatial information of patches (i.e., positional embedding) and this will provide more details on the locations of various features, as stated (Dosovitskiy et al., 2020):

$$z_0 = \left[x_p^1 E; x_p^2 E; \ldots..; x_p^N E\right] \oplus E_{pos} \tag{1}$$

where the projection of patches embedding is defined by $E \in R^{(P^2.C) \times D}$ and $E_{pos} \in R^{N \times D}$ illustrates the positional embedding matrix. L layers of multi-head self-attention (MSA) and multi-layer perceptron (MLP) blocks constitute the ViT encoder. As such, the feature representation of the lth layer is calculated as:

$$z_l' = MSA(LN(z_{l-1})) + z_{l-1} \tag{2}$$

$$z_l = MLP(LN(z_l')) + z_l' \tag{3}$$

where the layer normalization is expressed by $\mathbf{LN}(.)$ and the encoded image representation in lth layer is defined by z_l.

2.3.2. Decoder with HetConv layers and attention gates

Instead of using the classical CNN-based decoder network in the original TransU-Net segmentation algorithm, we utilized a decoder network that benefits from both additive attention gates and heterogeneous convolution (HetConv), as seen in Figs. 3 and 4. The concept of additive attention gates (AGs) were adopted from Oktay et al. (2018a) while the heterogeneous kernel convolutions were taken from Singh et al. (2019). The steps of the HetConv2D operations are explained in Algorithm 1 where $\oplus$ denotes pointwise addition operation.

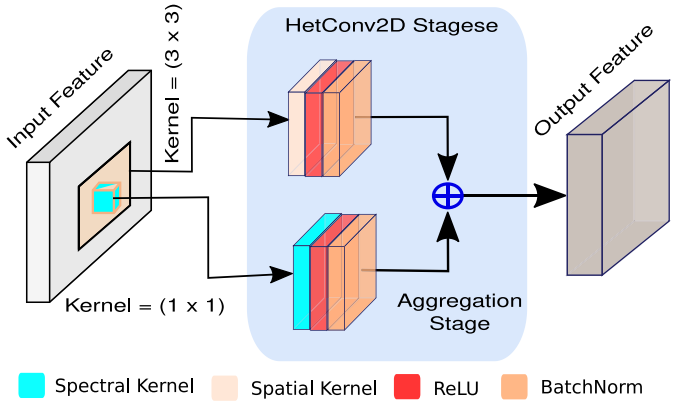


Fig. 4. Step of the heterogeneous kernel convolution, which combines both the features.

The feature map is gradually down-sampled in the architecture of a standard CNN model to acquire a reasonably large receptive field as well as the semantic contextual knowledge. Thus, low spatial level features can model the relationship between features and their respective location at the global scale. Nevertheless, reducing false-positive

Algorithm 1: HetConv2D Convolution.

Data: Input:X, Output:Y, $K^{(3,3)}$ and $K^{(1,1)}$
Function HetConv2D(X, $K^{(1,1)}_{point}$, $K^{(3,3)}_{depth}$):
  Y_{point} = BN(ReLU(Conv2D(X, $K^{(1,1)}$)))
  Y_{depth} = BN(ReLU(Conv2D(X, $K^{(3,3)}$)))
  $Y = Y_{point} \oplus Y_{depth}$
  return Y

estimation of small features with high shape variability remains challenging. To make the work easier to divide into different and future segmentation stages and increase accuracy, current segmentation designs rely on additional prior element localization principles. It was shown that the same goal could be accomplished by inserting AGs into a typical CNN model. This does not need training of multiple models or the addition of many extra parameters. In contrast with the localization concept employed in multi-stage CNNs, AGs restrict feature responses in the irrelevant regions without the need for cropping a region of interest (ROI) during feature extraction. As shown in Fig. 3, the salient areas of an image are identified by the attention coefficients, $\psi_i \in [0, 1]$, which then prune the feature responses to keep only the activations relevant to a given specific task. The element-wise multiplication of inputs and attention coefficients yields the result of AGs which is expressed as:

$$\hat{X}_i^l = X_i^l \otimes \psi_i^l \tag{4}$$

A single scalar attention value is generated for each pixel vector $X_i^l \in R^{B_l}$, where B_l refers to the quantity of feature maps present in the lth layer. A gating vector is applied to each pixel i is given a gating vector $g_i \in R^{B_g}$ to determine focus zones in an input image, as illustrated in Fig. 3. Thus, contextual knowledge is included in the gating vector to prune lower level feature maps. Additive attention is defined as:

$$s_{att}^l = \Omega^T(ReLU(\theta_x^T x_i^l + \theta_g^T g_i + b_g)) + b_\Omega \tag{5}$$

$$\psi_i^l = \sigma(s_{att}^l(x_i^l, g_i, W_{att})) \tag{6}$$

where W_{att} denotes the trainable parameters to compute the attention coefficient which are shared among the linear projection layers $\theta_g \in R^{B_g \times B_{int}}$, $\theta_x \in R^{B_l \times B_{int}}$, $\Omega \in R^{B_{int} \times 1}$ with the kernel of sizes $1 \times 1 \times 1$ and the bias vectors $b_g \in R^{B_{int}}$ and $b_\Omega \in R$. σ presents the *Sigmoid* activation function.

The decoder part of the network consists of four layers of UpSampling2D. The feature maps of lth transformer encoder layer X^l is passed through the UpSampling2D layer to match the shapes and the output of its corresponding AG layer, $\hat{X}_i^l$ are then concatenated which taken care in each level of the decoder network. The lth layer of the decoder network U_{dec}^l can be formulated as follows,

$$\tilde{X}^l = U_{dec}^l(\hat{X}_i^l, US(X^l), W_{het}) \tag{7}$$

where upsampling operation is denoted by US and W_{het} and $\tilde{X}$ represent the trainable parameter of the HetConv layer and output feature maps of the decoder network, respectively.

It should be noted that HetConv layers replace Conv2D layers, which are employed in the TransU-Net segmentation algorithm (Chen et al., 2021) to fuse the convolutional level of multiple receptive features. It is worth mentioning that the HetConv layers employ depth-wise convolutional groups with kernel sizes of (3×3) and a point-wise Conv2D with a kernel size of (1×1). It should be noted that the feature map of depth-wise convolutions is added to the point-wise convolution to produce the results of the HetConv functions, as illustrated in Fig. 4.

2.3.3. TransU-Net++ segmentation model

To increase the accurate prediction of deforestation mapping, an attention gates aided TransU-Net (**TransU-Net**++) segmentation algorithm is proposed, which utilizes the merits of the transformer and U-Net together with the attention gates to determine deforested regions using remote sensing images. Let us consider 4-band Amazon and Atlantic Forests as an example dataset, where the proposed network, **TransU-Net**++, takes the input images of sizes $128 \times 128 \times 4$. In the **TransU-Net**++, the initial feature extractor uses HetConv layers, and there are four groups of depth-wise Conv2D of sizes $128 \times 128 \times 32$ and a point-wise Conv2D of size $128 \times 128 \times 128$. The encoder network utilizes three max-pooling layers that down-samples the input image into an output map of size $16 \times 16 \times 128$. Afterward, the resulting feature map is passed to the transformer encoder layers. There are four blocks of transformer encoders in the developed **TransU-Net**++ segmentation algorithm where the final output map has a size of $8 \times 8 \times 44$, which is passed to a HetConv layer consisting of 4 groups of depth-wise Conv2D of sizes $8 \times 8 \times 32$ and a point-wise Conv2D of size $8 \times 8 \times 128$. The output feature map of the HetConv layer is sent to the decoder layer that uses an UpSampling2D layer as explained in the previous subsection, resulting in an output feature map of size $16 \times 16 \times 128$ and the attention gate with a feature map of size $16 \times 16 \times 64$. The HetConv layer and AG results are then concatenated as shown in Eq. (7), resulting in an output map of $16 \times 16 \times 192$. Afterward, the output map is sent to the second UpSampling2D layer, resulting in an output feature map of size $32 \times 32 \times 192$ and the attention gate with a feature map of size $32 \times 32 \times 16$. The HetConv layer and AG results are then concatenated, resulting in an output map of $32 \times 32 \times 208$. Then, the output map is sent to the third UpSampling2D layer, resulting in a feature map of size $64 \times 64 \times 208$ and the attention gate with a feature map of size $64 \times 64 \times 16$. The HetConv layer and AG results are concatenated, resulting in a feature map of $64 \times 64 \times 224$. Finally, the output map is sent to the last UpSampling2D layer, resulting in a feature map of size $128 \times 128 \times 224$. A 2D convolution with the kernel size of (1×1) and a sigmoid activation function are utilized at the last level of decoding to project the output map of the last HetConv layer into the targeted deforestation maps.

2.4. Comparison models

The developed model, TransU-Net++, is evaluated against several state-of-the-art segmentation models, including U-Net (Ronneberger et al., 2015), U-Net+++ (Huang et al., 2020), Attention U-Net (Oktay et al., 2018b), Swin U-Net (Cao et al., 2021), ResU-Net-a (Diakogiannis et al., 2020), SegNet (Badrinarayanan et al., 2017), ICNet (Zhao et al., 2018), ENet (Paszke et al., 2016), R2U-Net (Alom et al., 2018), Attention U-Net-2 (John and Zhang, 2022), and TransU-Net (Chen et al., 2021). The ResU-Net benefits from the architecture of the ResNet CNN model to improve the information propagation during training. The Swin U-Net utilizes the cutting-edge Swin Transformer algorithm, while the TransU-Net employs the first-generation vision Transformer to model the long-range dependencies in the segmentation framework. Both the Attention U-Net and Attention U-Net-2 take the advantage of the attention mechanism in their structures, while the R2U-Net utilizes a recurrent residual convolutional neural network (RRCNN) for improved representation learning. In addition, the U2-Net employs a two-level nested U-structure in its baseline structure. The SegNet was intended to be memory and computational time efficient during training. It also has a relatively limited amount of trainable variables. The encoder network's design is similar to the well-known VGG-16 CNN network. The ENet algorithm was primarily designed for applications with low latency operation for real-time pixel-wise semantic segmentation. To solve the problem of the major fraction of computation for pixel-wise label inference, ICNet segmentation model incorporates multi-resolution branches under proper label guidance.

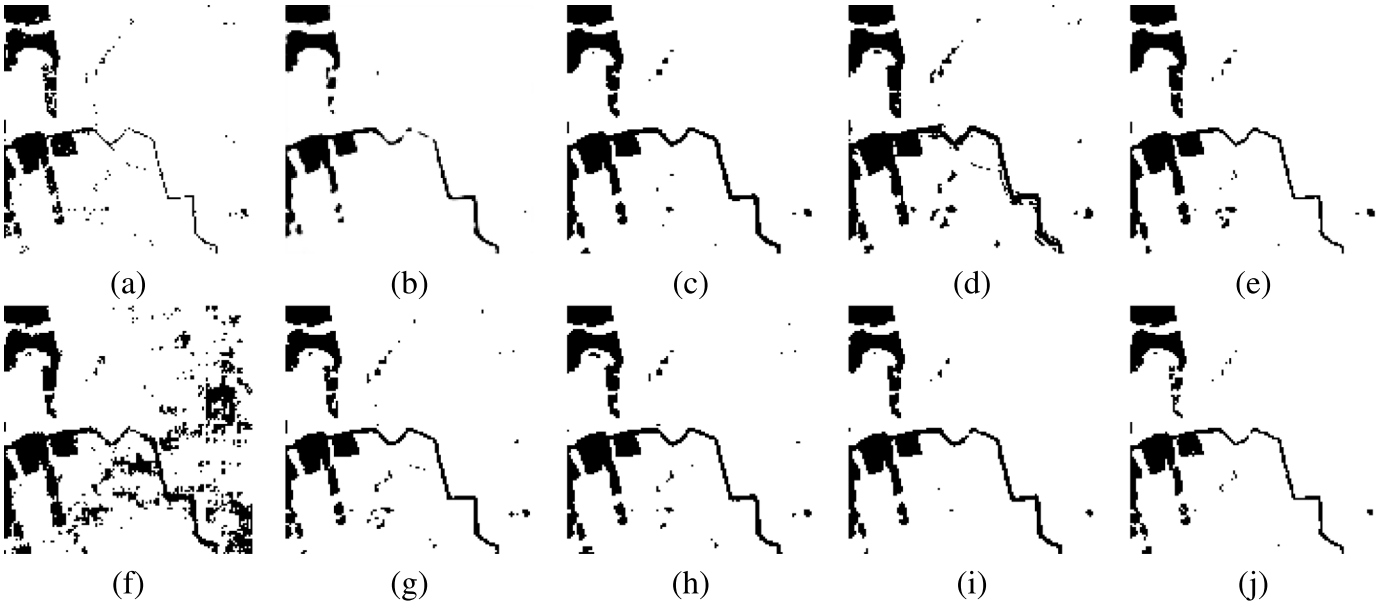


Fig. 5. Segmentation maps over 3-band Amazon Forest dataset obtained using (a) Ground Truth, (b) U-Net, (c) Attention U-Net, (d) R2U-Net, (e) ResU-Net (f) Swin U-Net, (g) U-Net+++, (h) Attention U-Net-2, (i) TransU-Net, and (j) TransU-Net++, respectively.

Table 1
Segmentation results of the Amazon 3-band Forest dataset in terms of Overall accuracy (%), F1-score (%), Precision (%), Recall (%), and Deforestation area, respectively.

Algorithm	OA	Precision	Recall	F1-score	Deforestation area (hectare)
UNet (Ronneberger et al., 2015)	86.61	79.24	97.50	87.38	1143.85
UNet+++ (Huang et al., 2020)	89.61	83.27	97.92	89.77	1148.89
R2UNet (Alom et al., 2018)	88.32	81.07	**98.55**	87.75	1156.25
AttUNet (Oktay et al., 2018b)	90.58	85.94	95.97	90.22	1126.01
AttUNet-2 (John and Zhang, 2022)	88.19	81.37	97.60	88.03	1145.54
SwinUNet (Cao et al., 2021)	89.20	86.10	92.28	89.11	1082.72
ResUNet-a (Diakogiannis et al., 2020)	89.34	84.06	95.85	88.68	1124.61
SegNet (Badrinarayanan et al., 2017)	79.01	80.98	73.25	77.53	859.43
ICNet (Zhao et al., 2018)	67.71	79.04	44.05	57.88	516.77
ENet (Paszke et al., 2016)	69.45	72.52	57.97	65.18	680.14
TransUNet (Chen et al., 2021)	88.61	82.98	95.78	88.55	1123.92
TransUNet++	**91.96**	**88.29**	95.88	**91.48**	1124.92
Ground Truth	–	–	–	–	1173.26

2.5. Accuracy evaluation

Segmentation results are assessed in terms of overall accuracy (OA), F-1 score, precision, and recall values.

$$OA = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

$$F - 1 = \frac{2 * TP}{2 * TP + FP + FN} \tag{11}$$

The TP, FP, and FN present true positive, false positive, and false negative values, respectively.

3. Results

3.1. 3-band Amazon forests segmentation results

As seen in Table 1, the TransU-Net++ shows the highest segmentation performance in terms of overall accuracy (91.96%), F1-score (91.48%), and precision (88.29%) as compared to the other segmentation models. The highest recall value was obtained by the R2U-Net algorithm (98.55%). Results indicate that the TransU-Net++ considerably improves the segmentation results of the TransU-Net algorithm by approximately 3%, 3%, and 6% in terms of F1-score, overall accuracy, and precision, respectively. In terms of visual interpretation, the best segmentation results with the least noisy maps were achieved by the Attention U-Net, R2U-Net, ResU-Net, and the TransU-Net+++, as shown in Fig. 5. The Swin U-Net and U-Net+++ showed much higher noises than other discussed algorithms. The least desirable segmentation results were obtained by SegNet, ICNet, and ENet segmentation models, as seen in Table 1. Moreover, as illustrated in Fig. 6, the R2U-Net segmentation model demonstrated the best deforestation zone mapping, although it resulted in the over-classification of deforested areas. On the other hand, the developed model of the TransU-Net++ segmentation algorithm shows much less confusion between deforested and forest areas as compared to the R2U-Net, while it considerably improves the results of the TransU-Net segmentation algorithm.

3.2. 4-band Amazon forests segmentation results

As seen in Table 2, the TransU-Net++ algorithm achieves the highest performance as compared to the other segmentation models in terms of overall accuracy (97.2%) and F1-score (97.18%). In addition,

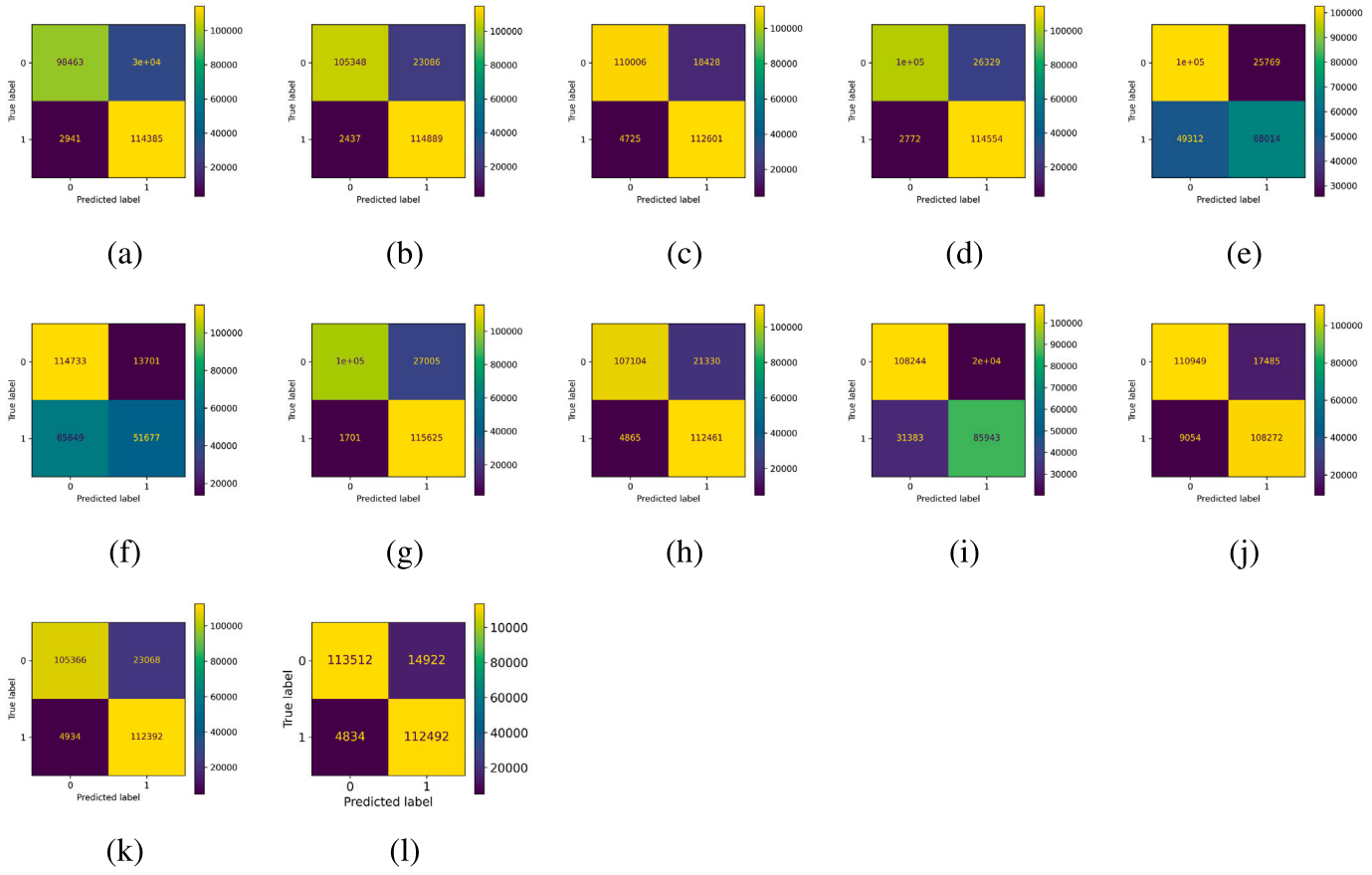


Fig. 6. Confusion matrices over 3-band Amazon Forest dataset obtained using (a) U-Net, (b) U-Net+++, (c) Attention U-Net, (d) Attention U-Net-2, (e) ENet, (f) ICNet, (g) R2U-Net, (h) ResU-Net, (i) SegNet, (j) Swin U-Net, (k) TransU-Net, and (l) TransU-Net++, respectively.

Table 2
Segmentation results of the Amazon 4-band Forest dataset in terms of Overall accuracy (%), F1-score (%), Precision (%), Recall (%), and Deforestation area, respectively.

Algorithm	OA	Precision	Recall	F1-score	Deforestation area (hectare)
UNet (Ronneberger et al., 2015)	96.52	94.14	99.23	96.59	1632.20
UNet+++ (Huang et al., 2020)	96.09	96.07	96.14	96.15	1581.32
R2UNet (Alom et al., 2018)	96.74	98.84	94.61	97.03	1556.17
AttUNet (Oktay et al., 2018b)	97	96.1	98	96.85	1611.96
AttUNet-2 (John and Zhang, 2022)	96.75	97.19	96.31	96.77	1584.08
SwinUNet (Cao et al., 2021)	92.94	88.7	98.49	92.74	1619.94
ResUNet-a (Diakogiannis et al., 2020)	92.88	**99.52**	86.24	93.61	1418.45
SegNet (Badrinarayanan et al., 2017)	92.66	92.78	92.59	92.47	1522.90
ICNet (Zhao et al., 2018)	89.53	91	87.84	89.36	1444.81
ENet (Paszke et al., 2016)	88.6	82.59	97.9	88.8	1610.74
TransUNet (Chen et al., 2021)	94.11	89.87	**99.48**	93.89	1636.20
TransUNet++	**97.2**	97.51	96.9	**97.18**	1593.89
Ground Truth	–	–	–	–	1644.83

the highest precision and recall values were obtained by the ResU-Net and TransU-Net algorithms with values of 99.52% and 99.48%, respectively. Moreover, the segmentation results of TransU-Net algorithm were considerably improved by the TransU-Net++ technique by approximately 3%, 3%, and 7% in terms of overall accuracy, F1-score, and precision, respectively. The least desirable segmentation results statically and visually were obtained by the SegNet, ENet, and ICNet models, as shown in Fig. 7. Although the TransU-Net algorithms resulted in the best recall value and better deforestation zone mapping as compared to the other segmentation algorithms, it demonstrated a much higher over-classification of the deforested area as compared to the developed TransU-Net++ algorithm, as seen in Fig. 8.

3.3. 4-band Atlantic forests segmentation results

As seen in Table 3, the TransU-Net++ algorithm obtained the highest segmentation performance compared to other segmentation models in terms of F1-score (90.57%), recall (93.96%), and overall accuracy (93.97%). The best segmentation results in terms of precision (89.53%) was achieved by Attention U-Net segmentation model. The TransU-Net++ segmentation technique substantially improved the results of the original TransU-Net technique by approximately 4%, 6%, and 16%, respectively, in terms of overall accuracy, F1-score, and recall, respectively, as shown in Table 3. In terms of visual interpretation, the best deforestation maps were obtained by the R2U-Net, Swin U-Net, U-Net+++, and TransU-Net++, as seen in Fig. 9. The worst deforestation

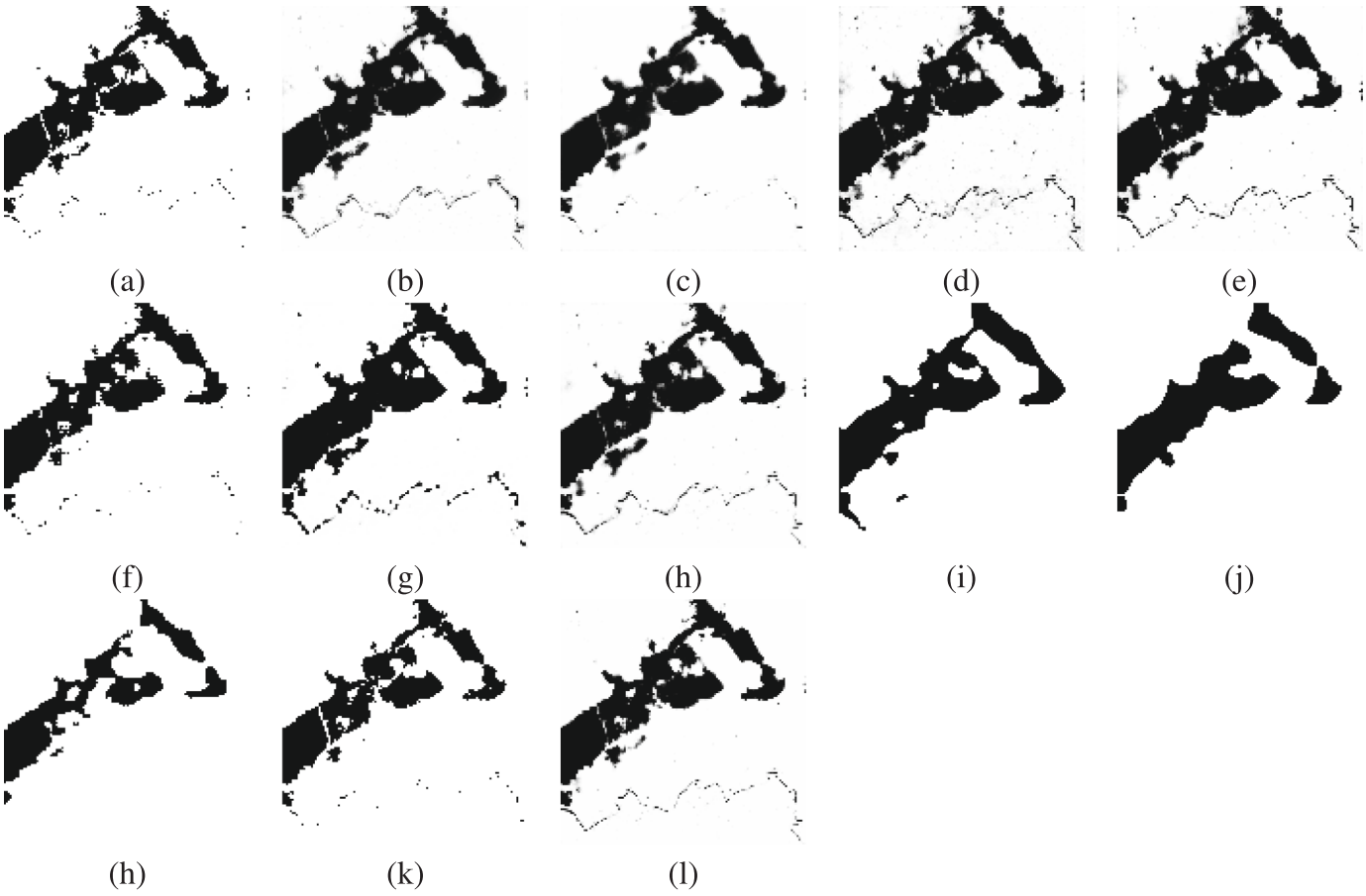


Fig. 7. Segmentation maps over 4-band Amazon Forest dataset obtained using (a) Ground Truth, (b) U-Net, (c) Attention U-Net, (d) R2U-Net, (e) ResU-Net (f) Swin U-Net, (g) U-Net+++, (h) Attention U-Net-2, (i) SegNet, (j) ICNet, (h) ENet, (k) TransU-Net, and (l) TransU-Net++, respectively.

Table 3
Segmentation results of the 4-band Atlantic Forest dataset in terms of Overall accuracy (%), F1-score (%), Precision (%), Recall (%), and Deforestation area, respectively.

Algorithm	OA	Precision	Recall	F1-score	Deforestation area (hectare)
UNet (Ronneberger et al., 2015)	88.84	76.42	93.44	84.05	965.13
UNet+++ (Huang et al., 2020)	87.4	87.91	69.6	77.92	718.87
R2UNet (Alom et al., 2018)	92.98	86.52	92.07	89.26	950.97
AttUNet (Oktay et al., 2018b)	91.83	**89.53**	83.89	87.97	866.59
AttUNet-2 (John and Zhang, 2022)	92.41	85.38	91.6	88.28	946.17
SwinUNet (Cao et al., 2021)	93.05	87.22	91.35	89.48	943.55
ResUNet-a (Diakogiannis et al., 2020)	92.99	87.78	90.34	89.32	933.18
SegNet (Badrinarayanan et al., 2017)	82.95	74.52	69.78	72.66	720.79
ICNet (Zhao et al., 2018)	78.52	68	60.18	65.53	621.73
ENet (Paszke et al., 2016)	83.43	73.63	73.87	74.45	763.05
TransUNet (Chen et al., 2021)	90.25	88.67	79.1	85.58	817.21
TransUNet++	**93.97**	87.76	**93.96**	**90.57**	970.54
Ground Truth	–	–	–	–	1032.92

maps were seen by SegNet and ICNet segmentation models. In addition, segmentation maps illustrated that the deforestation map of TransU-Net++ had a significant improvement as compared to the original TansU-Net segmentation algorithm, as seen in Fig. 9. As illustrated in Fig. 10, as compared to other segmentation algorithms, the developed TransU-Net++ model showed much less confusion for the deforestation zone mapping.

4. Discussion

4.1. Spatial transferability of segmentation algorithms

To evaluate the spatial transferability of the discussed segmentation algorithms, we trained them with the 4-band Amazon forests dataset and tested them with the 4-band Atlantic forests dataset. In terms

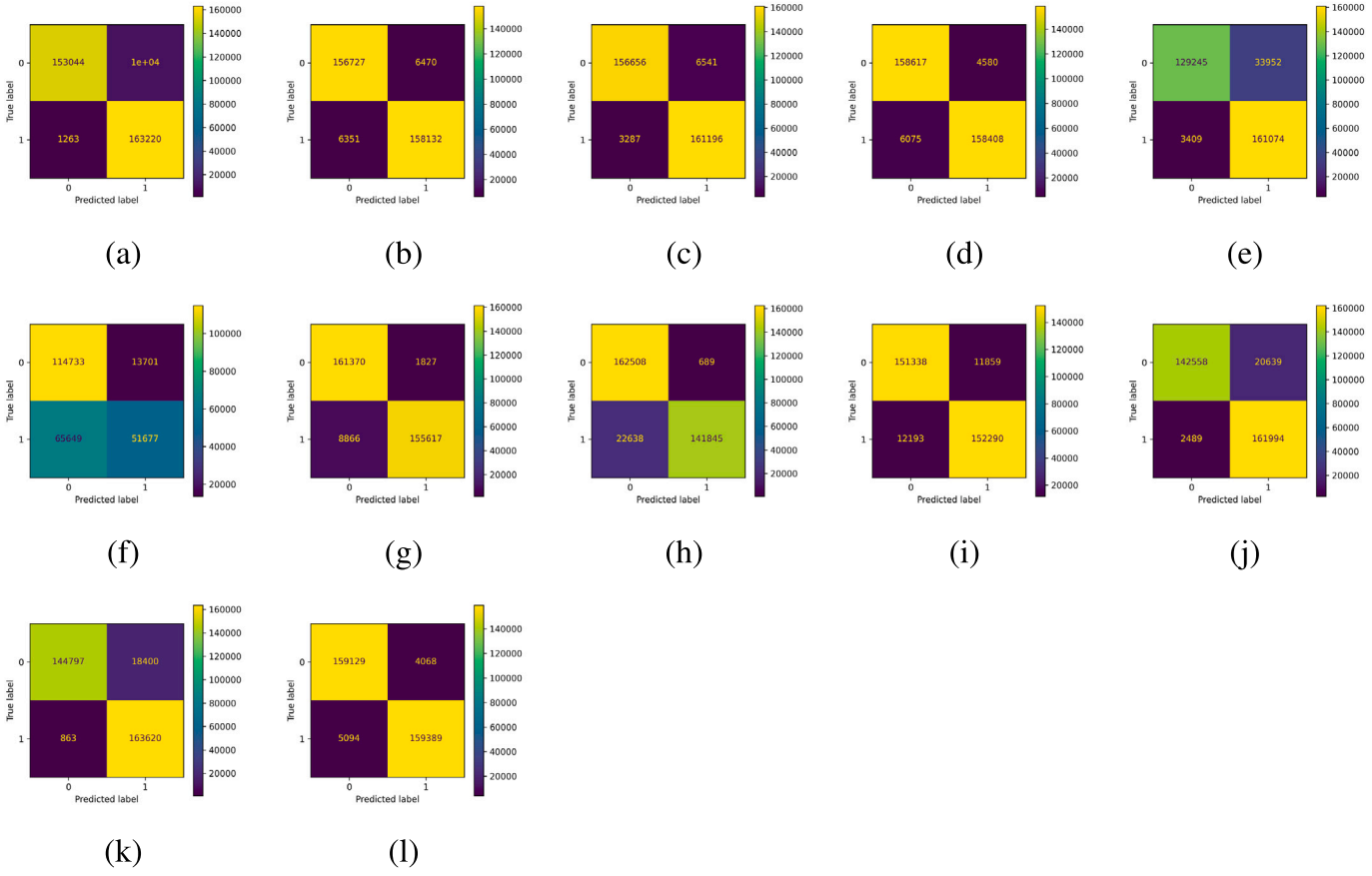


Fig. 8. Confusion matrices over 4-band Amazon Forest dataset obtained using (a) U-Net, (b) U-Net+++, (c) Attention U-Net, (d) Attention U-Net-2, (e) ENet, (f) ICNet, (g) R2U-Net, (h) ResU-Net, (i) SegNet, (j) Swin U-Net, (k) TransU-Net, and (l) TransU-Net++, respectively.

of precision, the TransU-Net++ considerably shows better statistical results as compared to the other segmentation algorithms, including ICNet, SegNet, ResU-Net, U-Net+++, Attention U-Net-2, U-Net, R2U-Net, Attention U-Net, TransU-Net, Swin U-Net, and ENet by approximately 23%, 19%, 9%, 8%, 8%, 5%, 4%, 4%, 3%, 3%, and 2%, respectively. Moreover, the TransU-Net++ algorithm obtained the highest segmentation performance as compared to other segmentation models in terms of precision (77.64%), F1-score (83.1%), and overall accuracy (88.21%), respectively. Moreover, the TransU-Net++ improves the segmentation results of the TransU-Net by about 1%, 2%, 2%, and 3%, respectively, in terms of recall, overall accuracy, F1-score, and precision, as seen in Table 4. The best segmentation results in term of recall was achieved by the Attention U-Net with a value of 94.18%. In terms of visual interpretation, the best deforestation maps were obtained by the Attention U-Net, R2U-Net, and TransU-Net++, while the worst visual results were obtained by SegNet, ICNet, and ENet segmentation models, as seen in Fig. 11. The worst deforestation maps were obtained by TransU-Net algorithm. Moreover, as illustrated in Fig. 12, there was less confusion between deforested and forest areas by the developed TransU-Net++ algorithm as compared to the TransU-Net segmentation model.

4.2. Area under the ROC curve

Moreover, Fig. 13 demonstrates the Area under the ROC Curve of different implemented segmentation models. Results proved the superiority of the developed TrasnU-Net++ model as compared to other segmentation algorithms obtaining the highest AUC values in all study areas, including Amazon-Atlantic (0.889), the 3-band Amazon (0.921), the 4-band Atlantic (0.94), and 4-band Amazon (0.972) datasets. Moreover, the TransU-Net++ algorithms improved the AUC value of the base TransU-Net model by approximately 1%, 3%, 3%, and 7% in the Amazon-Atlantic, 4-band Amazon, 3-band Amazon, and 4-band Atlantic datasets, respectively, as seen in Fig. 13.

4.3. Ablation study

We included the ablation study to better recognize the importance of the heterogeneous kernel convolutions and Attention gates in the developed segmentation model of the TransU-net++. As seen in Table 5, the inclusion of the HetConv and attention gates improved the segmentation results of the base TransU-Net model. The overall accuracy was improved by about 3%, 3%, and 4%, respectively, with the inclusion of the HetConv, attention gates, and both functions in the 3-band Amazon dataset. In the 4-band Amazon forest, the proposed functions of Hetconvs and attention gates increased the segmentation accuracy of the base TransU-Net model. For instance, the inclusion of the HetConv, attention gates, and both attention gates and HetConv functions considerably enhanced the segmentation results of the base TransU-Net model by approximately 5%, 7%, and 8%, respectively, in terms of precision statistical index, as seen in Table 6. Moreover, the inclusion of the attention gates, HetConv, and both attention gates and HetConv functions considerably enhanced the segmentation results of the base TransU-Net model by approximately 11%, 15%, and 16%, respectively, in terms of recall statistical index in the 4-band Atlantic dataset, as illustrated in Table 7. As illustrated in Table 8, the inclusion of the attention gates and Hetconvs enhanced the segmentation performance of the base TransU-Net algorithm. For example, the precision obtained by the TransU-Net model was increased by about 1% and 3%, respectively, by adding attention gates and both attention gates and HetConv in the Amazon to Atlantic forest dataset.

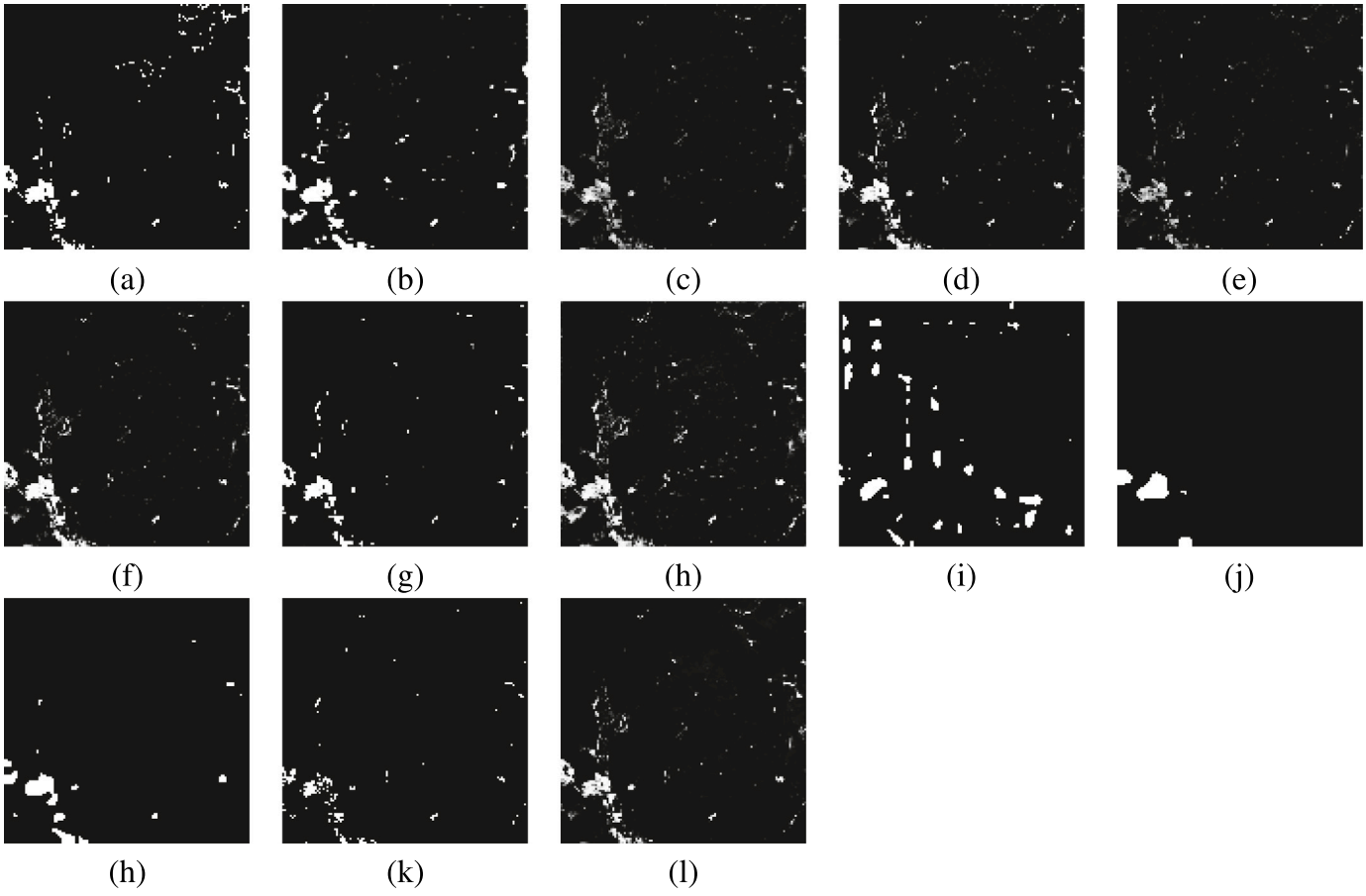


Fig. 9. Segmentation maps over 4-band Amazon Forest dataset obtained using (a) Ground Truth, (b) U-Net, (c) Attention U-Net, (d) R2U-Net, (e) ResU-Net (f) Swin U-Net, (g) U-Net+++, (h) Attention U-Net-2, (i) SegNet, (j) ICNet, (h) ENet, (k) TransU-Net, and (l) TransU-Net++, respectively.

Table 4
Segmentation results of the Amazon-Atlantic Forest dataset in terms of Overall accuracy (%), F1-score (%), Precision (%), Recall (%), and Deforestation area, respectively.

Algorithm	OA	Precision	Recall	F1-score	Deforestation area (hectare)
UNet (Ronneberger et al., 2015)	86.36	72.23	92.13	80.96	951.63
UNet+++ (Huang et al., 2020)	85.54	70.3	93.71	80.23	967.93
R2UNet (Alom et al., 2018)	87.2	73.22	93.65	81.71	967.35
AttUNet (Oktay et al., 2018b)	87.41	73.41	94.18	82.16	972.84
AttUNet-2 (John and Zhang, 2022)	85.53	70.16	94.11	79.44	972.06
SwinUNet (Cao et al., 2021)	87.3	73.71	92.84	81.93	959.00
ResUNet-a (Diakogiannis et al., 2020)	85.12	69.25	94.94	79.24	980.66
SegNet (Badrinarayanan et al., 2017)	78.16	61.9	79.9	69.01	825.29
ICNet (Zhao et al., 2018)	75.27	58.45	74.56	65.49	770.10
ENet (Paszke et al., 2016)	80.8	74.96	58.69	67.87	606.19
TransUNet (Chen et al., 2021)	86.87	73.93	90.14	81.4	930.58
TransUNet++	**88.21**	76.26	90.87	83.1	938.6
Ground Truth	–	–	–	–	1032.92

4.4. Feature maps visualization of AGs and CNN layers

To better visualize how the attention gates and convolutional layers generate intermediate features, here we present the feature maps of the first attention gate, last convolutional layer, and the output map of four randomly selected images resulting from the TransU-Net++ segmentation algorithm, as seen in Fig. 14. The results of the feature maps produced by the attention gates show how these function could precisely signify the deforested areas, resulting in a better model performance for forest lost mapping, as presented in Fig. 14. As discussed in previous subsections, the inclusion of the attention gates in the developed TransU-Net++ segmentation model significantly improved the results of the baseline TransU-Net model. For instance, in the Atlantic dataset, the TransU-Net++ segmentation technique significantly improved the segmentation results of the base TransU-Net model by about 4%, 6%, and 16%, respectively, in terms of statistical indices of overall accuracy, F1-score, and recall, respectively, as illustrated in Table 3. As such, achieved segmentation results signify and proves that the high importance of such concepts (i.e., attention gates) necessary in advanced and efficient segmentation algorithms.

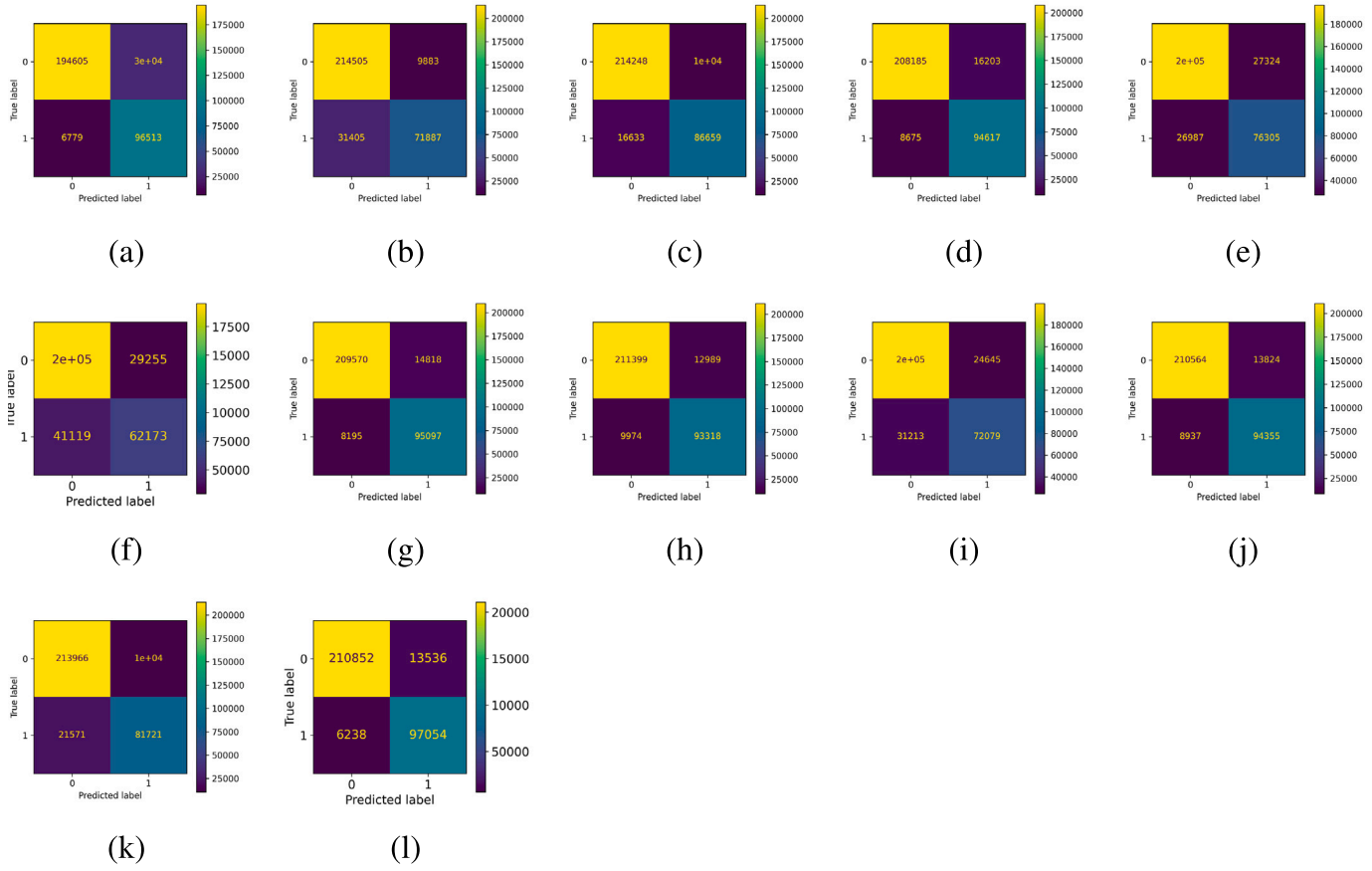


Fig. 10. Confusion matrices over 4-band Atlantic Forest dataset obtained using (a) U-Net, (b) U-Net+++, (c) Attention U-Net, (d) Attention U-Net-2, (e) ENet, (f) ICNet, (g) R2U-Net, (h) ResU-Net, (i) SegNet, (j) Swin U-Net, (k) TransU-Net, and (l) TransU-Net++, respectively.

Table 5
Segmentation results of the 3-band Amazon Forest dataset in terms of Overall accuracy (%), F1-score (%), Precision (%), and Recall (%), respectively.

Algorithm	OA	Precision	Recall	F1-score
TransU-Net (Chen et al., 2021)	88.61	82.98	95.78	88.55
TransU-Net-Het	91.50	87.33	96.15	90.97
TransU-Net-AGs	91.12	87.19	95.43	90.75
TransU-Net++	91.96	88.29	95.88	91.48

Table 6
Segmentation results of the 4-band Amazon Forest dataset in terms of Overall accuracy (%), F1-score (%), Precision (%), and Recall (%), respectively.

Algorithm	OA	Precision	Recall	F1-score
TransU-Net (Chen et al., 2021)	94.11	89.87	99.48	93.89
TransU-Net-Het	97.11	96.93	97.33	97.09
TransU-Net-AGs	96.79	95.09	98.69	96.58
TransU-Net++	97.20	97.51	96.9	97.18

Table 7
Segmentation results of the 4-band Atlantic Forest dataset in terms of Overall accuracy (%), F1-score (%), Precision (%), and Recall (%), respectively.

Algorithm	OA	Precision	Recall	F1-score
TransU-Net (Chen et al., 2021)	90.25	88.67	79.10	85.58
TransU-Net-Het	92.46	84.82	92.65	88.35
TransU-Net-AGs	92.14	86.45	89.03	88.08
TransU-Net++	93.97	87.76	93.96	90.57

Table 8
Segmentation results of the Amazon to Atlantic Forest dataset in terms of Overall accuracy (%), F1-score (%), Precision (%), and Recall (%), respectively.

Algorithm	OA	Precision	Recall	F1-score
TransU-Net (Chen et al., 2021)	86.87	73.93	90.14	81.40
TransU-Net-Het	83.88	67.09	95.90	78.25
TransU-Net-AGs	86.73	73.28	91.12	81.31
TransU-Net++	88.21	76.26	90.87	83.10

4.5. Computation cost

The computation costs of the developed segmentation models are compared in terms of time, as illustrated in Fig. 15. The highest computation cost was for training the segmentation models of ENet and Swin U-Net, as shown in Fig. 15 in all four datasets. The least required training time was seen for the model i.e., U-Net and U-Net+++. Moreover, the computation cost of the developed TransU-Net++ algorithm was slightly better as compared with the baseline TransU-Net segmentation model.

Although some of the segmentation models illustrated slightly better segmentation performance in some of the experimental datasets in terms of statistical indices, the proposed TransU-Net++ model illustrated superior segmentation results visually and statistically as compared to the base TransU-Net segmentation algorithm. Moreover,

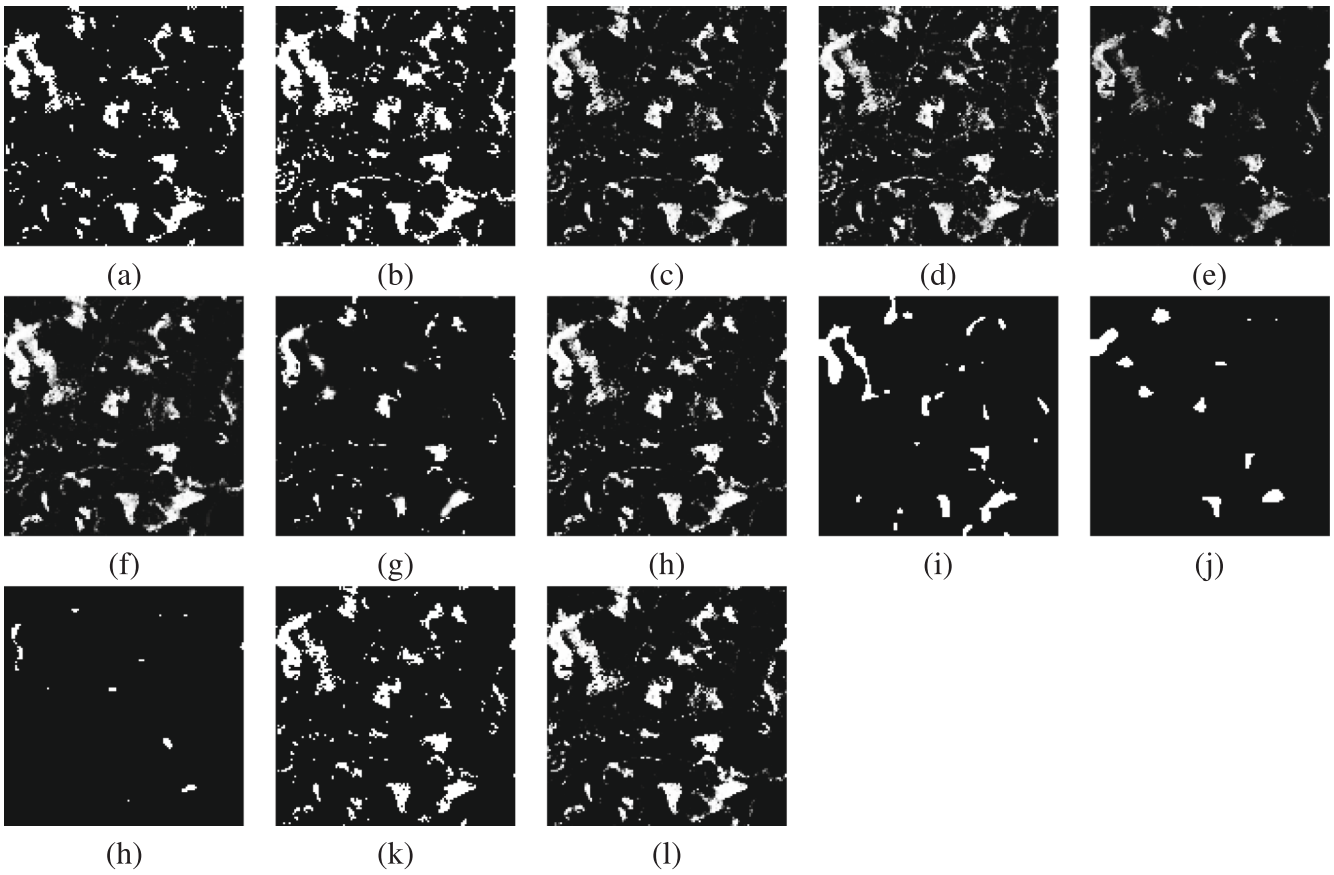


Fig. 11. Segmentation maps over 4-band Atlantic Forest dataset that was trained with the Amazon dataset (a) Ground Truth, (b) U-Net, (c) Attention U-Net, (d) R2U-Net, (e) ResU-Net (f) Swin U-Net, (g) U-Net+++, (h) Attention U-Net-2, (i) SegNet, (j) ICNet, (h) ENet, (k) TransU-Net, and (l) TransU-Net++, respectively.

based on the visual and statistical performances as discussed and presented in previous sections, overall the results produced by the TransU-Net algorithms proved to be more consistent as compared to the other segmentation models. For instance, results demonstrated the superiority of the developed TrasnU-Net++ model as compared to other segmentation algorithms obtaining the highest AUC values in all study areas, including Amazon-Atlantic (0.889), the 3-band Amazon (0.921), the 4-band Atlantic (0.94), and 4-band Amazon (0.972) datasets.

Though the developed segmentation model of TransU-Net++ was applied and evaluated for deforestation mapping, we do believe that the attention gates-aided version of the TransU-Net++ segmentation algorithm would show better and enhanced segmentation results as compared to the baseline TransU-Net model in other computer vision tasks (e.g., remote sensing segmentation, medical image segmentation and Smoke Semantic Segmentation, and much more). The different variants of TransU-Net proposed by many researchers (Zhang et al., 2021b; Yang and Mehrkanoon, 2022; He et al., 2022; Pan et al., 2023) etc. were due to the utilization of the widely used self-attention mechanism, becoming the successful segmentation framework. The reason behind this is because of the fine-grained long-range feature dependency captured through the self-attention in the transformer encoder block during training. In the future, the developed architecture could be evaluated in the medical image segmentation domain.

5. Conclusion

We present **TransU-Net**++, an enhanced attention gates-aided version of TransU-Net segmentation algorithm, for semantic segmentation. The proposed **TransU-Net**++ took advantage of the strengths and functionality of heterogeneous kernel convolution (HetConv), U-Net, attention gates, and vision transformers (ViTs). The developed model was evaluated for deforestation mapping using Sentinel-2 imagery. The obtained deforestation segmentation results proved the superiority of the **TransU-Net**++ algorithm as compared to several other cutting-edge CNN and vision-based segmentation models, including U-Net+++, Attention U-Net, Swin U-Net, ResU-Net-a, SegNet, ICNet, ENet, R2U-Net, and TransU-Net. Moreover, the developed **TransU-Net**++ model not only illustrated superior segmentation performance than other cutting-edge segmentation algorithms but also had a much better spatial information transferability. While trained on the Amazon forest and tested on the Atlantic forest, the **TransU-Net**++ algorithm achieved the highest segmentation accuracy (90.87%) over the other segmentation models in terms of recall, including ICNet, SegNet, ResU-Net, U-Net+++, Attention U-Net-2, U-Net, R2U-Net, Attention U-Net, TransU-Net, Swin U-Net, and ENet by approximately 23%, 19%, 9%, 8%, 8%, 5%, 4%, 4%, 3%, 3%, and 2% and substantially improved the segmentation results of the original TransU-Net segmentation technique by the margins of about 1%, 2%, 2%, and 3%, respectively, in terms of recall, overall accuracy, F1-score, and precision statistical indices. In addition, the results of ablation studies demonstrated the importance of the inclusion of the introduced concepts of the HetConvs and additive attention gates in the TransU-Net segmentation model. For instance, For instance, the inclusion of the HetConv, attention gates, and both attention gates and HetConv functions considerably enhanced the segmentation results of the base TransU-Net model by approximately 5%,

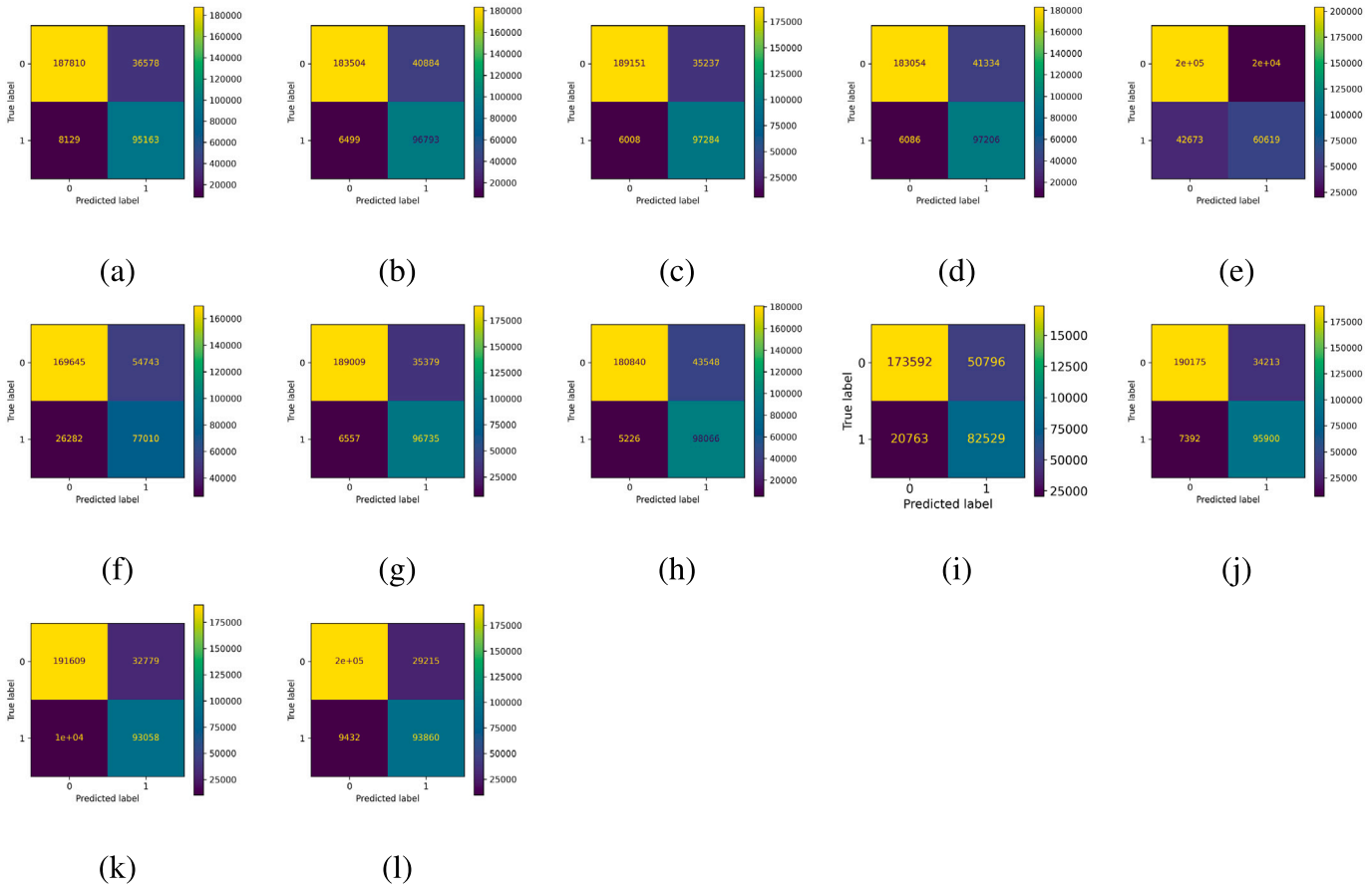


Fig. 12. Confusion matrices over 4-band Atlantic Forest dataset that was trained with the Amazon dataset using (a) U-Net, (b) U-Net+++, (c) Attention U-Net, (d) Attention U-Net-2, (e) ENet, (f) ICNet, (g) R2U-Net, (h) ResU-Net, (i) SegNet, (j) Swin U-Net, (k) TransU-Net, and (l) TransU-Net++, respectively.

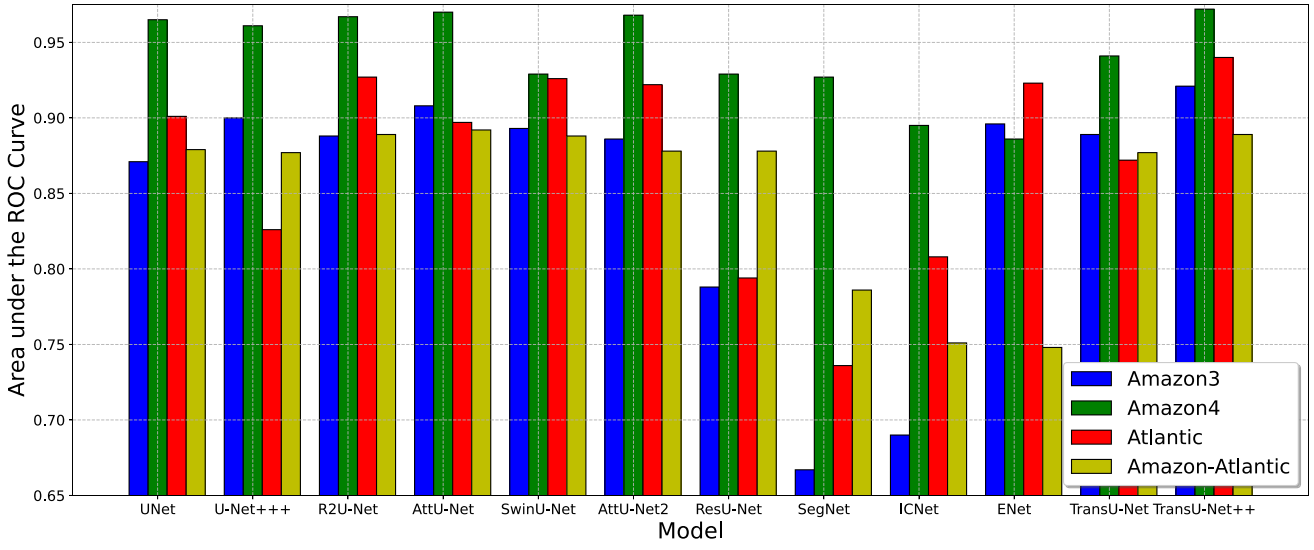


Fig. 13. Area under the ROC Curve of different segmentation models.

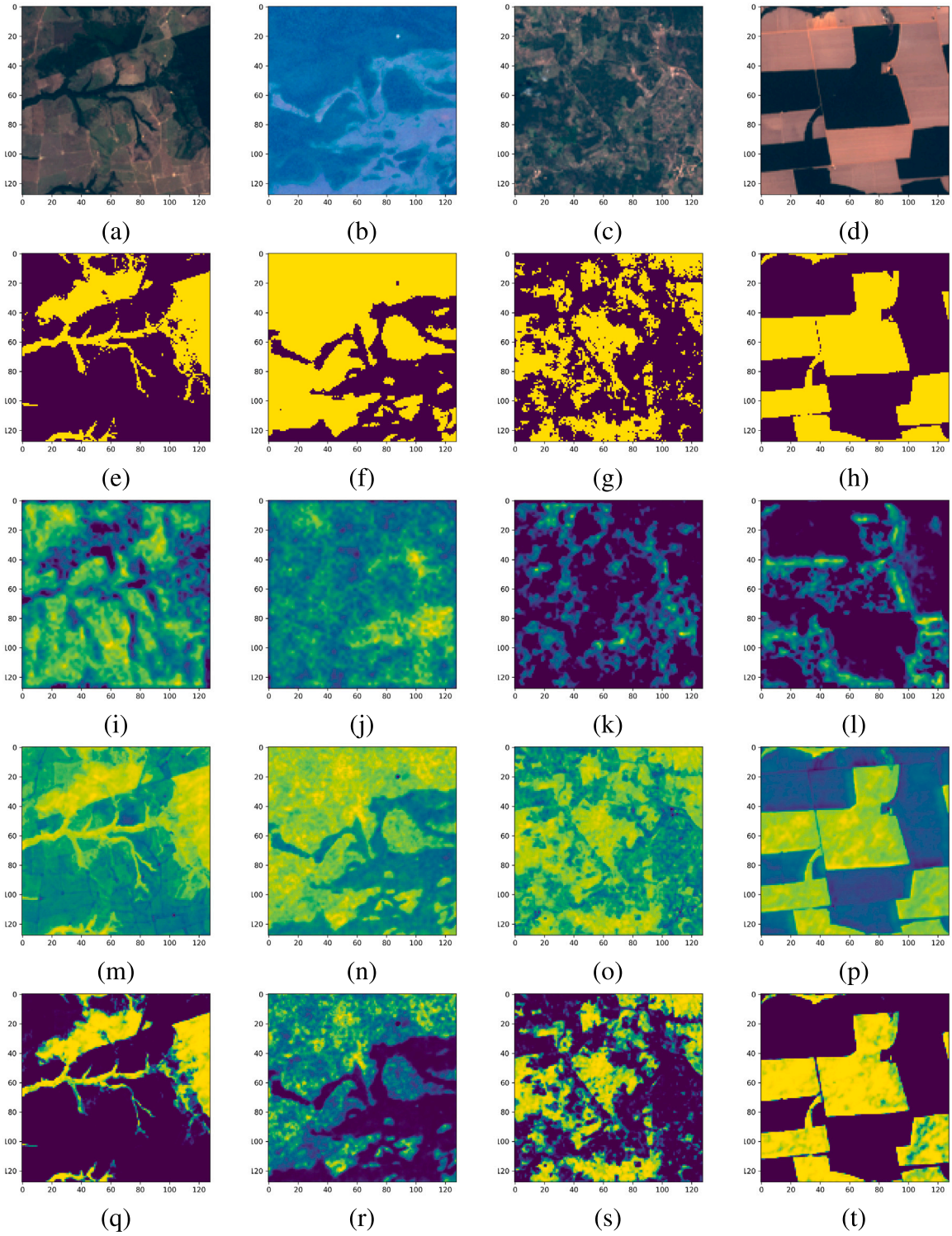


Fig. 14. Feature maps of the attention gates and the last layer of four deforestation images; (a)–(d) RGB images of four randomly selected images, (e)–(h) their respective deforestation ground truth maps, (i)–(l) their respective attention gate maps, (m)–(p) their feature maps of the last convolutional layer in the TransU-Net++ algorithm, and (q)–(t) their predicted deforestation maps using the TransU-Net++ algorithm.

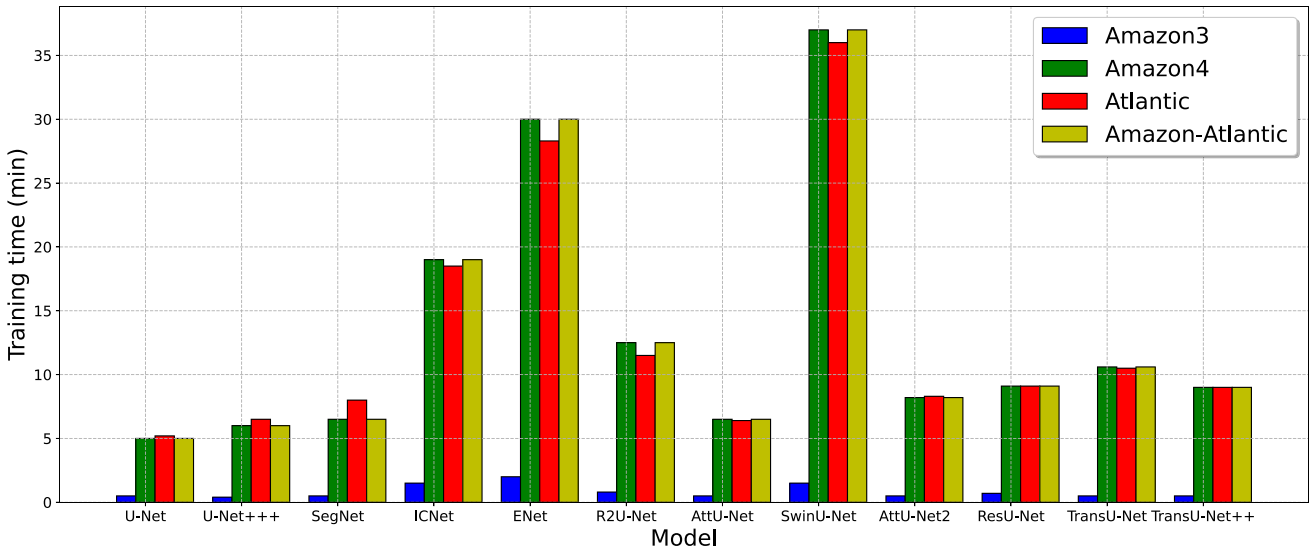


Fig. 15. A comparison of computation cost in term of running time over the different segmentation algorithms.

7%, and 8%, respectively, in terms of the precision statistical index in the 4-band Amazon Forest dataset.

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.

Data availability

Data will be made available on request.

Acknowledgments

This research is also funded by SERB, Govt. of India under Project Grant No. SRG/2022/001390.

References

Ahmad, M., Shabbir, S., Roy, S.K., Hong, D., Wu, X., Yao, J., Khan, A.M., Mazzara, M., Distefano, S., Chanussot, J., 2022. Hyperspectral image classification—Traditional to deep models: A survey for future prospects. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 15, 968–999. http://dx.doi.org/10.1109/JSTARS.2021.3133021.

Alom, M.Z., Hasan, M., Yakopcic, C., Taha, T.M., Asari, V.K., 2018. Recurrent residual convolutional neural network based on U-net (R2U-Net) for medical image segmentation. http://dx.doi.org/10.48550/ARXIV.1802.06955.

Alzu'bi, A., Alsmadi, L., 2022. Monitoring deforestation in Jordan using deep semantic segmentation with satellite imagery. Ecol. Inform. 70, 101745. http://dx.doi.org/10.1016/j.ecoinf.2022.101745.

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39 (12), 2481–2495. http://dx.doi.org/10.1109/TPAMI.2016.2644615.

Boers, N., Marwan, N., Barbosa, H.M.J., Kurths, J., 2017. A deforestation-induced tipping point for the South American monsoon system. Sci. Rep. 7 (1), 41489. http://dx.doi.org/10.1038/srep41489.

Bonan, G.B., 2008. Forests and climate change: forcings, feedbacks, and the climate benefits of forests. Sci. 320 (5882), 1444–1449.

Bragagnolo, L., da Silva, R., Grzybowski, J., 2021a. Amazon forest cover change mapping based on semantic segmentation by U-Nets. Ecol. Inform. 62, 101279. http://dx.doi.org/10.1016/j.ecoinf.2021.101279.

Bragagnolo, L., da Silva, R., Grzybowski, J., 2021b. Towards the automatic monitoring of deforestation in Brazilian rainforest. Ecol. Inform. 66, 101454. http://dx.doi.org/10.1016/j.ecoinf.2021.101454.

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M., 2021. Swin-unet: Unet-like pure transformer for medical image segmentation. arXiv Pr arXiv:2105.05537.

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y., 2021. Transunet: Transformers make strong encoders for medical image segmentation. arXiv Pr doi arXiv:2102.04306.

Diakogiannis, F.I., Waldner, F., Caccetta, P., Wu, C., 2020. ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. ISPRS J. Photogramm. 162, 94–114. http://dx.doi.org/10.1016/j.isprsjprs.2020.01.013.

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 16x16 words: Transformers for image recognition at scale. arXiv Pr arXiv:2010.11929.

Dutta, P., Sistu, G., Yogamani, S., Galván, E., McDonald, J., 2022. ViT-BEVSeg: A hierarchical transformer network for monocular birds-eye-view segmentation. In: 2022 International Joint Conference on Neural Networks. IJCNN, pp. 1–7. http://dx.doi.org/10.1109/IJCNN55064.2022.9891987.

Etteieb, S., Louhaichi, M., Kalaitzidis, C., Gitas, I.Z., 2013. Mediterranean forest mapping using hyper-spectral satellite imagery. Arab. J. Geosci. 6 (12), 5017–5032. http://dx.doi.org/10.1007/s12517-012-0748-6.

García-Ayllón, S., 2016. Rapid development as a factor of imbalance in urban growth of cities in Latin America: A perspective based on territorial indicators. Habitat Int. 58, 127–142. http://dx.doi.org/10.1016/j.habitatint.2016.10.005.

Ghamisi, P., Rasti, B., Yokoya, N., Wang, Q., Hofle, B., Bruzzone, L., Bovolo, F., Chi, M., Anders, K., Gloaguen, R., Atkinson, P.M., Benediktsson, J.A., 2019. Multisource and multitemporal data fusion in remote sensing: A comprehensive review of the state of the art. IEEE Trans. Geosci. Remote Sens. 7 (1), 6–39. http://dx.doi.org/10.1109/MGRS.2018.2890023.

Gong, P., Miller, J., Spanner, M., 1994. Forest canopy closure from classification and spectral unmixing of scene components-multisensor evaluation of an open canopy. IEEE Trans. Geosci. Remote Sens. 32 (5), 1067–1080. http://dx.doi.org/10.1109/36.312895.

Gulzar, Y., Khan, S.A., 2022. Skin Lesion segmentation based on vision transformers and convolutional neural networks-A comparative study. Appl. Sci. 12 (12), http://dx.doi.org/10.3390/app12125990, URL https://www.mdpi.com/2076-3417/12/12/5990.

Hamunyela, E., Reiche, J., Verbesselt, J., Herold, M., 2017. Using space-time features to improve detection of forest disturbances from landsat time series. Remote Sens. 9 (6), URL https://www.mdpi.com/2072-4292/9/6/515.

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., Kommareddy, A., Egorov, A., Chini, L., Justice, C.O., Townshend, J.R.G., 2013. High-resolution global maps of 21st-century forest cover change. Sci. 342 (6160), 850–853. http://dx.doi.org/10.1126/science.1244693.

Hao, S., Zhou, Y., Guo, Y., 2020. A brief survey on semantic segmentation with deep learning. Neurocomputing 406, 302–321. http://dx.doi.org/10.1016/j.neucom.2019.11.118.

He, J., Chen, J.-N., Liu, S., Kortylewski, A., Yang, C., Bai, Y., Wang, C., 2022. Transfg: A transformer architecture for fine-grained recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, no. 1. pp. 852–860.

Hesamian, M.H., Jia, W., He, X., Kennedy, P., 2019. Deep learning techniques for medical image segmentation: Achievements and challenges. J. Digit. Imaging 32 (4), 582–596. http://dx.doi.org/10.1007/s10278-019-00227-x.

Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., Han, X., Chen, Y.-W., Wu, J., 2020. Unet 3+: A full-scale connected unet for medical image segmentation.

In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 1055–1059.
Hubbell, S.P., He, F., Condit, R., de Água, L.B., Kellner, J., ter Steege, H., 2008. How many tree species are there in the Amazon and how many of them will go extinct? Proc. Natl. Acad. Sci. USA 105 (supplement_1), 11498–11504, https://www.pnas.org/doi/pdf/10.1073/pnas.0801915105.
Jamali, A., Mahdianpari, M., Brisco, B., Mao, D., Salehi, B., Mohammadimanesh, F., 2022a. 3DUNetGSFormer: A deep learning pipeline for complex wetland mapping using generative adversarial networks and swin transformer. Ecol. Inform. 72, 101904. http://dx.doi.org/10.1016/j.ecoinf.2022.101904.
Jamali, A., Mahdianpari, M., Mohammadimanesh, F., Homayouni, S., 2022b. A deep learning framework based on generative adversarial networks and vision transformer for complex wetland classification using limited training samples. Int. J. Appl. Earth Obs Geoinf 115, 103095. http://dx.doi.org/10.1016/j.jag.2022.103095.
Jamali, A., Roy, S.K., Bhattacharya, A., Ghamisi, P., 2023. Local window attention transformer for polarimetric SAR image classification. IEEE Geosci. Remote Sens. Lett. 1. http://dx.doi.org/10.1109/LGRS.2023.3239263.
John, D., Zhang, C., 2022. An attention-based U-net for detecting deforestation within satellite sensor imagery. Int. J. Appl. Earth Obs Geoinf. 107, 102685. http://dx.doi.org/10.1016/j.jag.2022.102685.
Kemker, R., Salvaggio, C., Kanan, C., 2018. Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning. ISPRS J. Photogramm. 145, 60–77. http://dx.doi.org/10.1016/j.isprsjprs.2018.04.014, Deep Learning RS Data.
Laurance, W.F., 2009. Conserving the hottest of the hotspots. Biol. Cons. 142 (6), 1137. http://dx.doi.org/10.1016/j.biocon.2008.10.011, Conservation Issues in the Brazilian Atlantic Forest.
Lausch, A., Erasmi, S., King, D.J., Magdon, P., Heurich, M., 2016. Understanding forest health with remote sensing-part I—A review of spectral traits, processes and remote-sensing characteristics. Remote Sens. 8 (12), URL https://www.mdpi.com/2072-4292/8/12/1029.
Lu, Z., Ding, C., Juefei-Xu, F., Boddeti, V.N., Wang, S., Yang, Y., 2023. Tformer: A transmission-friendly ViT model for IoT devices. IEEE Trans. Parallel Distrib. Syst. 34 (2), 598–610. http://dx.doi.org/10.1109/TPDS.2022.3222765.
Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., Johnson, B.A., 2019. Deep learning in remote sensing applications: A meta-analysis and review. ISPRS J. Photogramm. 152, 166–177. http://dx.doi.org/10.1016/j.isprsjprs.2019.04.015.
Malhi, Y., Roberts, J.T., Betts, R.A., Killeen, T.J., Li, W., Nobre, C.A., 2008. Climate change, deforestation, and the fate of the amazon. Sci. 319 (5860), 169–172, https://www.science.org/doi/pdf/10.1126/science.1146961.
Maslin, M., Malhi, Y., Phillips, O., Cowling, S., 2005. New views on an old forest: assessing the longevity, resilience and future of the Amazon rainforest. Trans. Inst. Br. Geogr. 30 (4), 477–499. http://dx.doi.org/10.1111/j.1475-5661.2005.00181.x.
Mikhaylov, A., Moiseev, N., Aleshin, K., Burkhardt, T., 2020. Global climate change and greenhouse effect. Entrepreneurship Sustain. Issues 7 (4), 2897–2913. http://dx.doi.org/10.9770/jesi.2020.7.4(21).
Müller, C., 2020. Brazil and the Amazon Rainforest : Deforestation, Biodiversity and Cooperation with the EU and International Forums. European Parliament, http://dx.doi.org/10.2861/520925.
Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M.C.H., Heinrich, M.P., Misawa, K., Mori, K., McDonagh, S.G., Hammerla, N.Y., Kainz, B., Glocker, B., Rueckert, D., 2018a. Attention U-Net: Learning where to look for the pancreas. CoRR abs/1804.03999 arXiv:1804.03999.
Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al., 2018b. Attention u-net: Learning where to look for the pancreas. arXiv Pr arXiv:1804.03999.
Pan, S., Liu, X., Xie, N., Chong, Y., 2023. EG-TransUNet: a transformer-based U-Net with enhanced and guided models for biomedical image segmentation. BMC Bioinformatics 24 (1), 1–22.
Paszke, A., Chaurasia, A., Kim, S., Culurciello, E., 2016. ENet: A deep neural network architecture for real-time semantic segmentation. arXiv:1606.02147.
Pires, G.F., Costa, M.H., 2013. Deforestation causes different subregional effects on the Amazon bioclimatic equilibrium. Geophys. Res. Lett. 40 (14), 3618–3623. http://dx.doi.org/10.1002/grl.50570.
Qi, X., Li, K., Liu, P., Zhou, X., Sun, M., 2020. Deep attention and multi-scale networks for accurate remote sensing image segmentation. IEEE Access 8, 146627–146639. http://dx.doi.org/10.1109/ACCESS.2020.3015587.
Rasti, B., Hong, D., Hang, R., Ghamisi, P., Kang, X., Chanussot, J., Benediktsson, J.A., 2020. Feature extraction for hyperspectral imagery: The evolution from shallow to deep: Overview and toolbox. IEEE Trans. Geosci. Remote Sens. 8 (4), 60–88. http://dx.doi.org/10.1109/MGRS.2020.2979764.
Rezende, C., Scarano, F., Assad, E., Joly, C., Metzger, J., Strassburg, B., Tabarelli, M., Fonseca, G., Mittermeier, R., 2018. From hotspot to hopespot: An opportunity for the Brazilian Atlantic Forest. Perspect. Ecol. Conserv. 16 (4), 208–214. http://dx.doi.org/10.1016/j.pecon.2018.10.002, URL https://www.sciencedirect.com/science/article/pii/S2530064418301317.
Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 234–241.
Roy, S.K., Chatterjee, S., Bhattacharyya, S., Chaudhuri, B.B., Platoš, J., 2020. Lightweight spectral–spatial squeeze-and- excitation residual bag-of-features learning for hyperspectral classification. IEEE Trans. Geosci. Remote Sens. 58 (8), 5277–5290.
Roy, S.K., Dubey, S.R., Chatterjee, S., Chaudhuri, B.B., 2020. FuSENet: fused squeeze-and-excitation network for spectral-spatial hyperspectral image classification. IET Image Process. 14 (8), 1653–1661.
Roy, S.K., Manna, S., Song, T., Bruzzone, L., 2021. Attention-based adaptive spectral-spatial kernel resnet for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. 59.
Samset, B.H., Fuglestvedt, J.S., Lund, M.T., 2020. Delayed emergence of a global temperature response after emission mitigation. Nature Commun. 11 (1), 3261. http://dx.doi.org/10.1038/s41467-020-17001-1.
Sandhya Devi, M.R., Vijay Kumar, V., Sivakumar, P., 2021. A review of image classification and object detection on machine learning and deep learning techniques. In: 2021 5th International Conference on Electronics, Communication and Aerospace Technology. ICECA, pp. 1–8. http://dx.doi.org/10.1109/ICECA52323.2021.9676141.
Scarano, F.R., Ceotto, P., Brazilian atlantic forest: impact, vulnerability, and adaptation to climate change. Biodiversity and Conservation 24 (9), 2319–2331. http://dx.doi.org/10.1007/s10531-015-0972-y.
Schulze, K., Malek, Žiga., Verburg, P.H., 2019. Towards better mapping of forest management patterns: A global allocation approach. Ecol. Manag. 432, 776–785. http://dx.doi.org/10.1016/j.foreco.2018.10.001.
Sexton, J.O., Song, X.-P., Feng, M., Noojipady, P., Anand, A., Huang, C., Kim, D.-H., Collins, K.M., Channan, S., DiMiceli, C., Townshend, J.R., 2013. Global, 30-m resolution continuous fields of tree cover: Landsat-based rescaling of MODIS vegetation continuous fields with lidar-based estimates of error. Int. J. Digit. Earth 6 (5), 427–448. http://dx.doi.org/10.1080/17538947.2013.786146.
Singh, P., Verma, V.K., Rai, P., Namboodiri, V.P., 2019. Hetconv: Heterogeneous kernel-based convolutions for deep cnns. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4835–4844.
Vanhala, P., Tamminen, P., Fritze, H., 2005. Relationship between basal soil respiration rate, tree stand and soil characteristics in boreal forests. Environ. Monit. Assess. 101 (1), 85–92. http://dx.doi.org/10.1007/s10661-005-9134-0.
Waldeland, A.U., Trier, Ø.D., Salberg, A.-B., 2022. Forest mapping and monitoring in africa using sentinel-2 data and deep learning. Int. J. Appl. Earth Obs Geoinf. 111, 102840. http://dx.doi.org/10.1016/j.jag.2022.102840.
Wang, L., Li, R., Zhang, C., Fang, S., Duan, C., Meng, X., Atkinson, P.M., 2022a. UNetFormer: A UNet-like transformer for efficient semantic segmentation of Remote Sensing urban scene imagery. ISPRS J. Photogramm. 190, 196–214. http://dx.doi.org/10.1016/j.isprsjprs.2022.06.008.
Wang, L., Li, R., Zhang, C., Fang, S., Duan, C., Meng, X., Atkinson, P.M., 2022b. UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. ISPRS J. Photogramm. Remote Sens. 190, 196–214.
Wang, W., Tang, C., Wang, X., Zheng, B., 2022c. A ViT-based multiscale feature fusion approach for remote sensing image segmentation. IEEE Geosci. Remote Sens. Lett. 19, 1–5. http://dx.doi.org/10.1109/LGRS.2022.3187135.
Yang, Y., Mehrkanoon, S., 2022. Aa-transunet: Attention augmented transunet for nowcasting tasks. In: 2022 International Joint Conference on Neural Networks. IJCNN, IEEE, pp. 01–08.
Yin, H., Khamzina, A., Pflugmacher, D., Martius, C., 2017. Forest cover mapping in post-soviet central Asia using multi-resolution Remote Sensing imagery. Sci. Rep. 7 (1), 1375. http://dx.doi.org/10.1038/s41598-017-01582-x.
Yuan, F., Zhang, Z., Fang, Z., 2023. An effective CNN and Transformer complementary network for medical image segmentation. Pattern Recognit. 136, 109228.
Zhang, W., Li, J., Hua, Z., 2021a. Attention-based tri-UNet for remote sensing image pan-sharpening. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 14, 3719–3732. http://dx.doi.org/10.1109/JSTARS.2021.3068274.
Zhang, Y., Liu, H., Hu, Q., 2021b. Transfuse: Fusing transformers and cnns for medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24. Springer, pp. 14–24.
Zhao, H., Qi, X., Shen, X., Shi, J., Jia, J., 2018. Icnet for real-time semantic segmentation on high-resolution images. In: Proceedings of the European Conference on Computer Vision. ECCV.
Zhao, Z.-Q., Zheng, P., Xu, S.-T., Wu, X., 2019. Object detection with deep learning: A review. IEEE Trans. Neural Netw. 30 (11), 3212–3232. http://dx.doi.org/10.1109/TNNLS.2018.2876865.