Neural Computing and Applications
https://doi.org/10.1007/s00521-024-09743-6

ORIGINAL ARTICLE

Enhancing the ability of convolutional neural networks for remote sensing image segmentation using transformers

Mohammad Barr[1]

Received: 27 February 2023 / Accepted: 25 March 2024

Abstract
The segmentation of remote sensing images has emerged as a compelling undertaking in computer vision owing to its use in the development of several applications. The U-Net style has been extensively utilized in many picture segmentation applications, yielding remarkable achievements. Nevertheless, the U-Net has several constraints in the context of remote sensing picture segmentation, mostly stemming from the limited scope of the convolution kernels. The transformer is a deep learning model specifically developed for sequence-to-sequence translation. It incorporates a self-attention mechanism to efficiently process many inputs, selectively retaining the relevant information and discarding the irrelevant inputs by adjusting the weights. However, it highlights a constraint in the localization capability caused by the absence of fundamental characteristics. This work presents a novel approach called U-Net–transformer, which combines the U-Net and transformer models for the purpose of remote sensing picture segmentation. The suggested solution surpasses individual models, such as U-Net and transformers, by combining and leveraging their characteristics. Initially, the transformer obtains the overall context by encoding tokenized picture patches derived from the feature maps of the convolutional neural network (CNN). Next, the encoded feature maps undergo upsampling through a decoder and are then merged with the high-resolution feature maps of the CNN model. This enables the localization to be more accurate. The transformer serves as an unconventional encoder for segmenting remote sensing images. It enhances the U-Net model by capturing localized spatial data, hence improving the capacity to capture intricate details. The U-Net–transformer, as suggested, has demonstrated exceptional performance in remote sensing picture segmentation across many benchmark datasets. The given findings demonstrated the efficacy of integrating the U-Net and transformer model for the purpose of segmenting remote sensing images.

Keywords Aerial images · Segmentation · Convolutional neural networks · Transformers

1 Introduction

Recently, remote sensing data have been widely used for earth exploration and surveillance. For a small portion of the earth, satellite vision sensors capture remote sensing images. The geographic information system generally uses remote sensing images [1] for cartography purposes. For low-resolution images, the pixel intensity is enough for manual classification, but with recent high-resolution images, it is a hard challenge to classify different regions and time-consuming. Besides, high-resolution images have much more information and details that cannot be extracted manually such as color, shapes, texture, density, and structure. The geographic information system provides acceptable results on low-resolution images but struggles with high-resolution images that require more powerful techniques. Image segmentation techniques [2] based on artificial intelligence were proposed as a solution.

Image segmentation tends to present objects in the image based on their classification density at the pixel level. It is used to enhance the representation of the image and make it easier to analyze. There are two types of image segmentation: semantic segmentation [3] and instance segmentation [4]. Semantic segmentation associates each image pixel with a particular class, but instance

✉ Mohammad Barr
mohammed.barr@nbu.edu.sa

[1] Department of Electrical Engineering, College of Engineering, Northern Border University, Arar, Saudi Arabia

Published online: 26 April 2024

segmentation treats multiple objects of the same object as different instances. Semantic segmentation is generally used because the main task is to classify regions in the remote sensing image. Region classification can be used for many applications such as land use and land cover detection [35, 36], weather forecasting, forest fire detection, and water resources studies.

A powerful semantic segmentation approach must segment different regions with high precision. Many approaches have been proposed for semantic segmentation, such as semantic texton forest [5] and random forest [6]. However, most of those algorithms are based on handcrafted features that rely on the expertise of the feature extraction engineer. This type of algorithm may provide satisfactory results for particular cases but not for other cases. Besides, many challenges will be encountered in the segmentation task, such as image noise, non-uniform intensity, occlusion, and deformation. So, algorithms based on handcrafted features are not a good solution for segmenting complex, high-resolution remote sensing images. Self-learning algorithms, known as deep learning [7], are considered the best candidates for the studied task due to their ability to learn more comprehensive features directly from the input data.

Deep learning models are based on deep neural networks such as convolutional neural networks (CNN) and recurrent neural networks (RNN). Those networks have been widely adopted in real-world applications such as object detection [8], scene recognition [9], traffic sign detection [10], and pedestrian detection [11]. The performance of deep learning models comes from the depth of the learned features, the ability to learn different features, and the fusion of those features to provide predictions. A wide range of applications have proven the superiority of deep learning models compared to other models based on handcrafted features.

Motivated by the massive success of deep learning models for many applications, including image segmentation, we proposed using a CNN model inspired by the U-Net model [12]. However, CNN models present many limitations when modeling explicit long-range relations. As the main convolution operation is locally performed, the model's performance is degraded when dealing with high-resolution remote sensing images. To avoid such limitations, recent works have proposed the adoption of self-attention modules. Transformers [13], the recent advance in self-attention modules, were designed for sequence-to-sequence prediction based on replacing dispense convolution operations with attention modules. Unlike regular CNN models, transformers have high transferability for downstream applications and high performance in global contexts. Transformers have been recently adopted for image classification tasks, and superior results were achieved [14].

In this work, we propose studying transformers' impact on remote sensing image segmentation. However, after intensive experimentation, we figured out that using the native transformers for image segmentation does not work well and low accuracy was achieved. This was caused by the processing methodology of the transformers which take a 1D input sequence and focus on modeling the global context at each stage. So, the low-resolution features will lack rich localization information, which will affect the resulting segmentation map. Furthermore, this is caused by the non-recovery of those features by a simple upsampling process. Subsequently, CNN models have great power in extracting rich semantic features that can be quickly recovered through deconvolution.

Considering the advantages and limitations of the CNN model and transformers, we proposed to enhance the CNN model with transformers for semantic segmentation of remote sensing images. The transformers were used to compensate for global context modeling loss, and the CNN model was used to compensate for feature resolution loss. The proposed U-Net–transformer is a hybrid model composed of CNN and transformers. The CNN model inspired by the U-Net style first employed the transformers to generate the self-attention features and then upsampled those features to be combined with high-resolution features of the CNN model residually connected from the transformers to collect high-precision localization information. The proposed design showed the efficiency of combining the CNN model and the transformers for semantic segmentation of remote sensing images.

The combination of U-Net architecture with transformer for remote sensing image segmentation brings together the strengths of both architectures, contributing to improved performance in handling complex and high-resolution remote sensing images. U-Net is known for its effectiveness in capturing local features and details through its encoder–decoder structure. On the other hand, transformers excel at capturing global context and long-range dependencies. By combining U-Net with transformers, the model can leverage both local and global information, enhancing its ability to understand and segment remote sensing images. Transformers incorporate attention mechanisms that allow the model to focus on relevant parts of the input. This is particularly beneficial for remote sensing images where specific regions of interest may vary in size and scale. The attention mechanisms enable the model to adaptively allocate resources to important features during segmentation. Remote sensing images often have high spatial resolution, leading to a vast amount of data. U-Net with transformers can efficiently handle this high-resolution data by capturing both local details and global context. The

attention mechanisms in transformers can be particularly useful in managing large image sizes and extracting relevant information.

Remote sensing images may exhibit diverse characteristics, including varying land cover types, environmental conditions, and seasonal changes. The combined model benefits from the adaptability of transformers to diverse input patterns, making it more robust and capable of generalizing well across different remote sensing scenarios. The combined U-Net and transformer architecture may reduce the reliance on extensive preprocessing steps. The model can learn to extract meaningful features directly from raw or minimally processed remote sensing imagery, simplifying the overall workflow and potentially improving efficiency. The transformer's self-attention mechanism allows the model to capture intricate relationships between pixels, leading to an improved semantic understanding of the remote sensing scene. This is valuable for accurately delineating boundaries between different land cover classes.

In summary, the integration of U-Net with transformer models brings a synergy of local and global context information, attention mechanisms, and adaptability to diverse image characteristics. This combination contributes to more effective and accurate segmentation of remote sensing images, addressing the challenges posed by high-resolution and complex environmental conditions.

Extensive experiments demonstrate that the proposed hybrid model presents an efficient technique for semantic segmentation of remote sensing images compared to regular CNN models, even those enhanced by self-attention modules different from the transformer's design. Besides, we proved that intensive incorporation of low-level features results in better segmentation performance. Reported empirical results demonstrated the efficacity and robustness of the U-Net–transformer model against state-of-the-art models for semantic segmentation of remote sensing images.

The main contributions to this work are the following:

(1) Developing a semantic segmentation system for remote sensing images based on deep learning techniques.
(2) We propose a hybrid model by combining a CNN model with skip connections and transformers in the U-Net style.
(3) We are evaluating the performance of the proposed hybrid model on two different datasets.

The organizer of this work is the following: Sect. 2 will represent related works. In Sect. 3, the proposed approach will be presented and detailed. Experiments and results will be presented and discussed in Sect. 4. Finally, in Sect. 5, conclusions and future works will be provided.

2 Related works

Semantic segmentation of remote sensing images is an active research field due to its importance for many applications such as urban planning, crop planning, disaster management, and environmental preservation. As a result, many works in the literature have been proposed to achieve reliable results for the semantic segmentation task.

Qi et al. [15] proposed a deep learning-based model for remote sensing image segmentation. The proposed model, named ATD-LinkNet, was developed by upgrading the building blocks of D-LinkNet [16] with an attention module. The main idea of the D-LinkNet was to build an encoder–decoder network based on dilated convolution layers [17]. Besides, a pre-trained encoder was used to achieve high performance. So, the proposed model took advantage of the base model and applied some enhancements by integrating an attention module. The new building block collects features from different scales and fuses them, enabling it to use the rich spatial and semantic information of remote sensing images.

A novel technique called deep adaptation-based change detection technique (DACDT) [38] is suggested for processing optical and SAR images using an image translation process-oriented approach. A refined U-Net + + model is introduced to enhance the overall and localized effects of the photos. Furthermore, a multi-scale loss function is employed to evaluate the characteristics of several dimensions. The final change maps are constructed by transferring the characteristics of optical pictures to the SAR images in order to provide more accurate change analysis. The suggested approach's prediction performance is assessed on four distinct datasets: Gloucester I, Shuguang Village, Gloucester-II, and California. The estimated outputs determine the predictive performance of the suggested solution, with accuracy rates of 98.67, 99.77, 97.68, and 98.87%, respectively.

A modified U-Net model named TL-DenseUNet [18] was proposed for semantic segmentation of remote sensing images. The proposed model was composed of two parts. The first part is the encoder based on the DenseNet model [19]. The DenseNet model was pretrained on the ImageNet dataset [20], and the transfer learning technique was applied to use the model for extracting semantic features from remote sensing images. The second part is the decoder, which adopts the connection methodology of the building blocks of the DenseNet model. This part merges the features of multiscale layers. The decoder network was frozen, and only the decoder network was fine-tuned using the remote sensing images from the Sparse Representation and Intelligent Analysis competition [18]. The proposed TL-DenseUNet has achieved an overall accuracy of

72.01%. The reported results show that the proposed model improved against other models, but those results are insufficient for a high-performance semantic segmentation technique.

Liang et al. [21] proposed the combination of a Swin transformer [31] and a CNN model for remote sensing image segmentation. The proposed method was based on extracting coarse-grained and fine-grained feature representations at various semantic scales using a staged model. Furthermore, for maximum benefit from the extracted features, a fusion module based on the self-attention mechanism from the transformer was proposed to fuse features collected at different stages. The evaluation of the proposed method on the Vaihingen and Potsdam datasets proved its efficiency compared to baseline models.

As it is known that deep learning models require a large amount of training data and the available remote sensing images are few, it was essential to handle this limitation. It was proposed in [22] to use game data to train deep learning models to segment real-world remote sensing data. The proposed method was based on the combination of a cycle-generative adversarial model (Cycle GAN) [23] and the ResNet model [24] for the segmentation task. The cycle GAN was used for image-to-image translation, which maps synthetic images from the game to the remote sensing images. The ResNet model was used for the segmentation of the remote sensing images. Both models were trained end-to-end using only synthetic images without real remote sensing images. The evaluation of the proposed method proved its efficiency with an intersection over union (IoU) of 0.521. The proposed method is up-and-coming and can be improved to achieve reliable performance for real-world applications.

An enhanced CNN model was proposed in [25] for remote sensing image segmentation. The main idea was to integrate additional modules to improve the performance of a CNN model. An adaptive multi-scale module (AMSM) and an adaptive fuse module (AFM) were proposed. Despite the high performance of CNN models for solving computer vision tasks, they still suffer from some challenges when dealing with high-resolution remote sensing images, such as difficulty in collecting relevant features and direct integration of external components by the traditional encoder–decoder models. The AMSM was used to fuse input channels by configuring the structures using different void rates to handle the mentioned challenges. Then, it generates weights based on the analysis of image content. Finally, the AFM was applied to deep feature maps to obtain their consequences for filtering the noise in high-resolution feature maps. The integration of those modules has enhanced the overall accuracy of the model.

The transformers were applied for remote sensing image segmentation by Xu et al. [26]. A Swin transformer backbone enhanced the original transformer model, and implicit and explicit edge enhancement techniques were applied to handle the edge detection problem. Four versions of the proposed transformers were proposed. The efficient L model has achieved the best overall accuracy on the Potsdam and Vaihingen datasets.

Li et al. [27] proposed a multitask semantic boundary awareness network to segment remote sensing images. The main concept of the proposed model named SBANet was to integrate boundary attention modules for multitask learning through adaptive weight. First, the boundary attention module was used to collect relevant features of the object boundary through hierarchical feature aggregation in a bottom-up manner. Then, the multitask learning fuses a boundary loss with the semantic loss. So, the boundary extraction module and the semantic segmentation will work jointly. The evaluation of the SBANet on two different datasets proved its efficiency for remote sensing image segmentation.

A transformer network called SwinSUNet [37] was developed, which utilizes a Siamese U-shaped topology to effectively address change detection difficulties. The SwinSUNet architecture has encoder, fusion, and decoder components, all of which utilize Swin transformer blocks as fundamental elements. The encoder utilizes a Siamese structure that is built around a hierarchical Swin transformer. This allows the encoder to efficiently handle bitemporal pictures by processing them in parallel and extracting their multiscale characteristics. Fusion primarily facilitates the merging process of the bitemporal characteristics produced by the encoder. The decoder, similar to the encoder, utilizes a hierarchical Swin transformer. In contrast to the encoder, the decoder employs an upsampling and merging block and Swin transformer blocks to restore the specifics of the change information. The encoder employs patch merging and Swin transformer blocks to provide very efficient semantic features. Upon completing the sequential procedure of these three modules, SwinSUNet will generate the change maps. Our research involved conducting costly tests on four change detection datasets.

After a deep study of existing works, we figured out that most works are underperforming and that more performances must be achieved. In this work, we proposed the integration of the transformers with a CNN model to achieve high performances that allow for building a reliable and robust remote sensing image segmentation system. The next section provides a detailed description of the proposed approach.

3 Proposed approach

This section introduces the proposed approach for segmenting remote sensing images. Considering an input image, the main goal is to predict the label map at the pixel level. Generally, to perform this task, encoder–decoder models are used to encode the input image into high compressed features and then decode those features back to the input resolution. The main idea of this work was to embed the attention mechanism in the decoder by using the transformers. The U-Net style was adapted for building the CNN model used for the segmentation. First, we will explain how to integrate the transformers into the encoder, and then, the full model will be represented.

Based on transformer topology, the input sequence must be tokenized by generating flattened 2D patches with a size of p × p. The total number of patches N for each image can be calculated as (1).

$$N = \frac{hw}{p^2} \tag{1}$$

where p is the patch size, h is the height of the image, and w is its width.

To integrate the transformers into the encoder, the vectorized patches were mapped using a trainable linear projection into a latent D-dimensional embedding space. For encoding the spatial information of the patch, specific position embeddings were learned and then added to the patch embeddings to remember the positional information. So, the encoded patch can be represented as (2).

$$z_0 = \left[x_p^1\mathrm{E}; x_p^2\mathrm{E}; \ldots; x_p^N\mathrm{E}\right] + \mathrm{E}_{\mathrm{pos}} \tag{2}$$

where E is the embeddings projection of the patch, and $\mathrm{E}_{\mathrm{pos}}$ is the embedding position.

As an encoder, transformers are composed of L alternating layers of Multi-head Self-Attention (MSA) modules and Multi-Layer Perceptron (MLP). So, the output of the l-th layer can be computed as (3) and (4).

$$z'_l = MSA(\mathrm{LN}(z_{l-1})) + z_{l-1} \tag{3}$$

$$z_l = MPL(\mathrm{LN}(z'_l)) + z'_l \tag{4}$$

$LN(.)$ is the layer normalization function, and z_L is the encoded output. The MLP comprises two layers with Gaussian Error Linear Unit (GELU) nonlinear activation function. The GELU can be computed as (5) where $\sigma = 1$.

$$GELU(x) = x\sigma(1.702x) \tag{5}$$

To generate the final segmentation map, it is most common to upsample the encoder features to the resolution of the input image and then predict the labels for each pixel. Using transformers, recovering the spatial order can be performed by reshaping the encoder features from $\frac{hw}{p^2}$ to $\frac{h}{p} \times \frac{w}{p}$. The number of channels must be reduced to the number of classes using 1 × 1 convolution layers. In the end, feature maps are bilinearly upsampled to h × w resolution, generating predictions.

Using transformers as encoders and then using naive upsampling for generating the final segmentation map can achieve reasonable accuracy, but it is still under the expected. Besides, this strategy has many drawbacks, such as losing low-level details due to the big difference between the resolution of the encoded feature maps and the input resolution. This problem results in a degradation in the model's performance, especially for applications that require a sharp segmentation map. Therefore, the proposed U-Net–transformer model was proposed to compensate for performance degradation. It has a hybrid structure composed of a CNN model, transformers for feature encoding, and a cascaded upsampling technique to retain localization information. Figure 1 presents an overview of the proposed model.

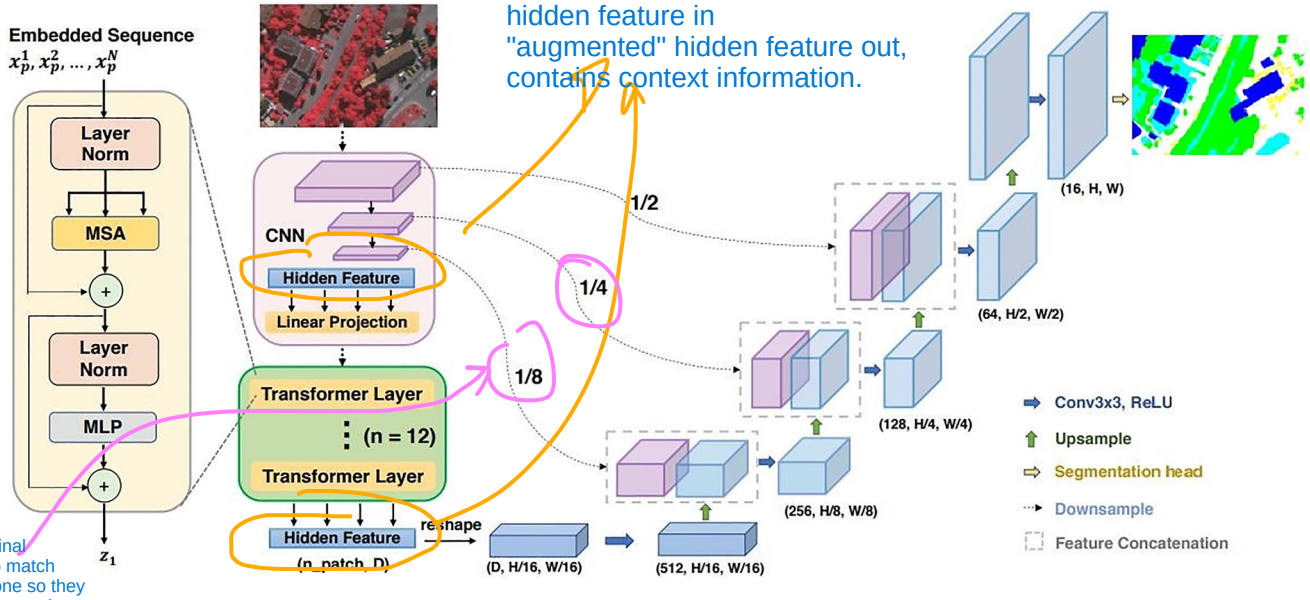
Instead of using the transformers for features encoding, we proposed to combine them with a CNN model. First, the CNN model was used to extract features and reduce the resolution of the feature maps. Second, transformers take 1 × 1 patches extracted from the feature maps generated by the CNN model rather than using the image directly as input. The proposed structure of the encoding stage has two main advantages, which are:

(1) Allows leveraging intermediate high-resolution feature maps from the CNN in the decoding path.
(2) The combination of the CNN and transformers allows taking advantage of the power of both models.

As a final stage for the segmentation task, we proposed a cascaded upsampling technique. It is composed of multiple upsampling steps to convert the compressed feature maps to the original resolution of the input. In the first stage, the decoder compressed the input from $h \times w$ size to $\frac{h}{p} \times \frac{w}{p}$ size. So, the proposed multi-step upsampling technique was applied to the features maps with a resolution of $\frac{h}{p} \times \frac{w}{p}$ until resolving $h \times w$. Each upsampling step comprises a convolution layer with a 3 × 3 kernel and a nonlinear activation layer based on the rectified linear units (ReLU).

In summary, combining the encoder based on the CNN model and transformers with the cascaded upsampling technique follows the U-Net style. This style allows the aggregation of features at different stages through skip connections. That was very important for the reconstruction of the output and the generation of the segmentation mask. The proposed approach combines the CNN model and transformers to work as an encoder. Both models were used to take advantage of each other. Using the native

as far as I understood here, they don't do transpose convolution. They do standard convolution than upsampling, which is a little bit different I think. We can try to apply transpose convolution instead and see if the performance changes.



hidden feature in "augmented" hidden feature out, contains context information.

resize the original feature map to match the upscaled one so they can be concatenated. This idea comes from the U-Net paper.

Fig. 1 Structure of the proposed U-Net–transformer model

transformers for image segmentation does not work well due to the processing methodology of the transformers that take a 1D input sequence and focus on modeling the global context at each stage. So, the low-resolution features will lack rich localization information, which will affect the resulting segmentation map. The non-recovery of those features causes this by a simple upsampling process. Subsequently, CNN models have great power in extracting rich semantic features that can be quickly recovered through deconvolution.

kinda old hardware, we might even try to do stuff locally.

4 Experiments and results

4.1 Experimental environment and evaluation

All the experiments of this work were performed using a desktop equipped with an Intel i7 CPU, 32 GB of RAM, and Nvidia GTX 960 GPU. The models were developed based on the TensorFlow deep learning framework with the support of CUDA and cuDNN accelerations. In addition, the open cv was used for data load and visualization.

The proposed approach was evaluated using two public datasets. First, the 2D ISPRS Potsdam dataset [28] was used. It is composed of high-resolution images with a resolution of 5cm. The images consist of four near-infrared (NIR) channels and red, blue, and green orthorectified imagery collected using aircraft. The dataset provides the digital surface model corresponding to the images. Besides, it provides ground-truth segmentation masks. The dataset provides 6 different classes, including surfaces, buildings, trees, low vegetation, cars, and unknown objects. The 2D ISPRS Potsdam dataset provides a total of 38 images, where only 24 of them are labeled and used for training and evaluation.

The images have a resolution of 6000 × 6000 with four channels mentioned above. To use the images for training and validation, slices with a size of 256 × 256 were extracted and stored in separate batches. Second, the DeepGlobe Road Extraction dataset [32] was used. The dataset consists of RGB images with 50 cm per pixel resolution. It was collected from the south of Asia by DigitalGlobe's satellite. The resolution of the images is 1024 × 1024 pixels. The dataset consists of a total of 8579 images divided into three sets. The first set contains 6226 images for training, the second set contains 1243 images for validation, and the last set contains 1110 images for testing. The dataset presents a binary segmentation task intending to extract roads from satellite images. The output segmentation mask is in grayscale, white represents roads, and black represents the background. Both datasets were used to evaluate the proposed model's performance for segmenting remote sensing images.

roads only

For training, we applied different data augmentation techniques such as rotation, translation, and flipping to make the model more robust against real-world challenges. For both CNN and transformers models, pretrained weights on ImageNet [20] were used for initialization. The input resolution was fixed to 512 × 512 for training and testing, and the patch size was fixed to 16. The Adam optimizer was used for training with an initial learning rate of 0.01 and a weight decay of 1e-4. Due to the limitations of the available GPU, we fixed the batch size to 4. The training was performed for 100 k iterations of the DeepGlobe Road

100

Extraction dataset and 70k for the 2D ISPRS Potsdam dataset. We used the efficient net model [29] for the CNN model, and for the transformers, we used the ViT model [30].

As evaluation metrics, the overall accuracy and the mIoU were used. The evaluation metrics were chosen based on benchmark datasets and the existing state-of-the-art works. The experiments were conducted on both datasets and compared to state-of-the-art works. We performed many configurations to prove the proposed hybrid model's robustness. First, we used an encoder composed of a CNN model and transformers and a decoder based on naive upsampling. Second, we used the proposed cascaded upsampling technique instead of naive upsampling. Finally, to make a fair comparison against existing works, we selected those evaluated on the same dataset and reproduced their work in our experimental environments. Table 1 presents the overall accuracy and mIoU on the 2D ISPRS Potsdam dataset.

The experiments on the 2D ISPRS Potsdam dataset have been performed to compare the performances of the proposed model against state-of-the-art models. We conducted many experiments on the proposed model with a different configuration. To prove the efficiency of the proposed model comparison against state-of-the-art works was performed. Since the official evaluation metric for the 2D ISPRS Potsdam dataset was used for comparison purposes in addition to the mIoU metric. As it is shown in Table 1, the U-Net–transformer with cascaded upsampling achieved the best overall accuracy and the best mIoU. It has an overall accuracy higher than the ATD-LinkNet [15] with 1.43% and more than 18% higher than U-Net [12]. The visual output of the U-Net–transformer and state-of-the-art models is presented in Fig. 2.

To further improve the efficiency of the proposed model, we evaluate it using the DeepGlobe Road Extraction dataset. The dataset presents a binary segmentation problem where only the roads class is considered. Furthermore, this dataset considers the mIoU as an official evaluation metric. The experimental results proved the efficiency of the U-Net–transformer for road extraction tasks. Table 2 provides the achieved results in comparison with state-of-the-art models.

As shown in Table 2, the U-Net–transformer with cascaded upsampling performs better than the U-Net–transformer with naive upsampling. Both models have higher mIoU compared to state-of-the-art models. In addition, it was proved that the proposed model has a higher capability for segmentation. The visual output of the U-Net–transformer with naïve upsampling and cascaded upsampling compared to the output of state-of-the-art models is presented in Fig. 3.

4.2 Ablation study

To thoroughly prove the efficiency of the proposed model, we conducted a series of experiments in different settings. Then, we performed many ablation studies to validate the performance of our model. As configurations, we manipulated the number of skip connections with upsampling stage, resolution of the input images, length of the input sequence of the transformer, patch size, and model scaling as we presented in the proposed approach section that integrates skip connections between the encoder and decoder which enhances the collection of finer segmentation details by transmitting low-level spatial features to the output. The ablation study aimed to evaluate the impact of the number of skip connections on the model's performance. Starting with integrating one skip connection at the ¼ resolution then at ¼ and ½ until reaching three skip connections at ¼, ½, and 1/8 resolutions scales. We noticed that adding more skip connections led to higher performances. So, we adopted the three skip connections in our model. It was important to mention that the segmentation of small objects is highly improved compared to large objects. This achievement proved the efficacity of integrating the CNN model and the transformers in the decoder, the cascaded upsampling use for the decoder, and the importance of the skip connection at different resolution scales. Table 3 presents the variation of the

Table 1 Achieved results in terms of overall accuracy and mIoU on the 2D ISPRS Potsdam dataset

Model	Overall accuracy (%)	mIoU (%)
ATD-LinkNet [15]	89	74.31
TL-DenseNet [18]	72.01	43.08
CGFCN [22]	73.46	52.18
U-Net [12]	72.35	53.77
MFNet [25]	90.47	77.05
Efficient L [26]	89.44	82.68
SBANet [27]	90.59	82.77
U-Net–transformer with naive upsampling	90.43	83.34
U-Net–transformer with cascaded upsampling	91.02	85.41

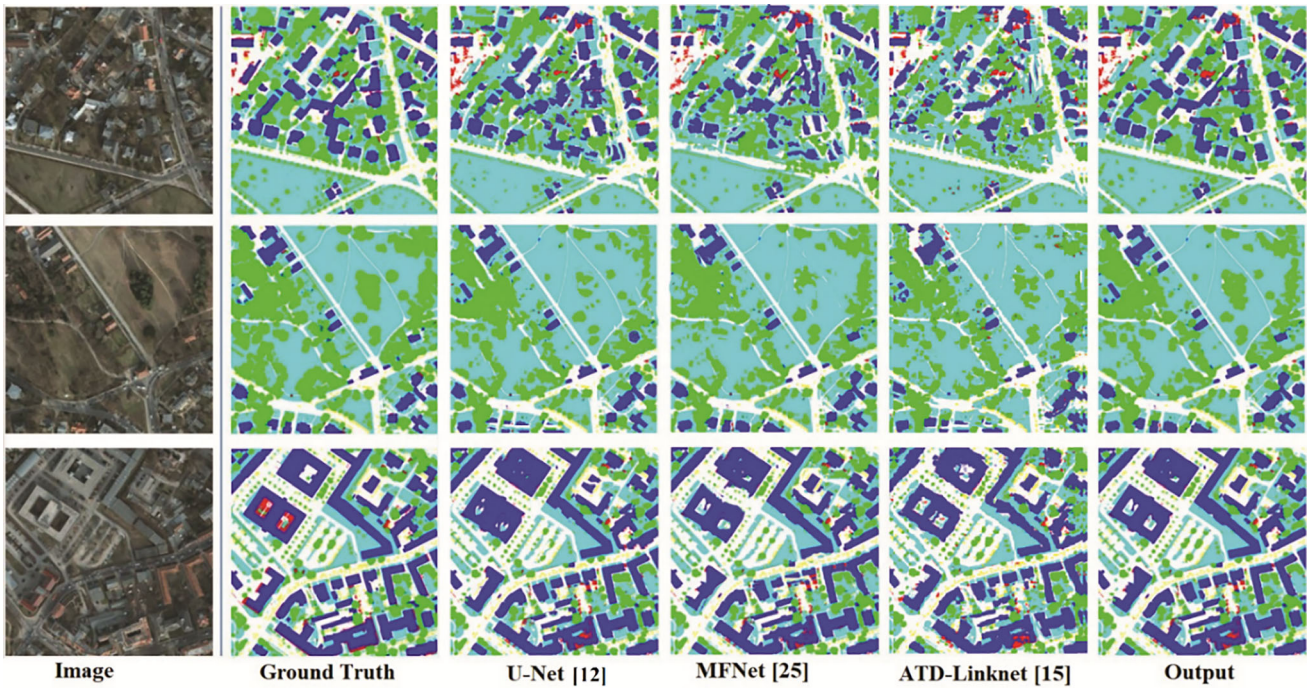


Fig. 2 Visual output of the model compared to state-of-the-art works. The proposed model presents more accuracy in crowded space as shown in the first row. The segmentation is more sharp and clearly shows different regions

Table 2 Achieved results in terms of overall accuracy and mIoU on the DeepGlobe Road Extraction dataset

Model	Overall accuracy (%)	mIoU (%)
ATD-LinkNet [15]	84.54	62.68
TL-DenseNet [18]	70.93	41.09
CGFCN [22]	79.76	50.93
U-Net [12]	69.42	49.67
MFNet [25]	88.34	74.56
Efficient L [26]	84.11	78.45
SBANet [27]	88.93	79.42
U-Net–transformer with naive upsampling	88.73	80.22
U-Net–transformer with cascaded upsampling	89.09	83.87

performances in relation to the number of skip connections on the 2D ISPRS Potsdam dataset.

The default resolution of the images in both datasets is 1024 × 1024. This ablation study evaluated the impact of the size of the input image on model performances. We started by using 512 × 512 image size while fixing the patch size to 16; then, we moved to the original image size of 1024 × 1024. Using larger images generates five times larger sequences as input to the transformer. The increasing image size has improved the overall accuracy by 1.88% but strongly affected the processing speed. So, we ignored the achieved improvements in performance to balance the accuracy and processing speed. Table 4 presents the achieved results using different input sizes.

Furthermore, we performed an ablation study on the impact of the patch size. We variate the patch from 8 to 16 and then 32. After analyzing the results, we observed that the best accuracy was achieved with a smaller patch size of 8. However, using a very small patch size degrades the processing speed and requires a large storage memory. So, we considered a patch size of 16 to balance the performances between the accuracy, processing speed, and required memory. Table 5 summarizes the achieved performances in relation to the patch size.

Besides, an ablation study was performed to evaluate the impact of the model scale. For this purpose, we proposed two configurations. The first model has 3072 hidden layers, an MLP size of 768, several heads of 12, and the size of D is 12 and the second one has double each component. The experiments proved that the larger model has better accuracy and lower processing speed, while the light model has lower accuracy and better processing speed. Table 6

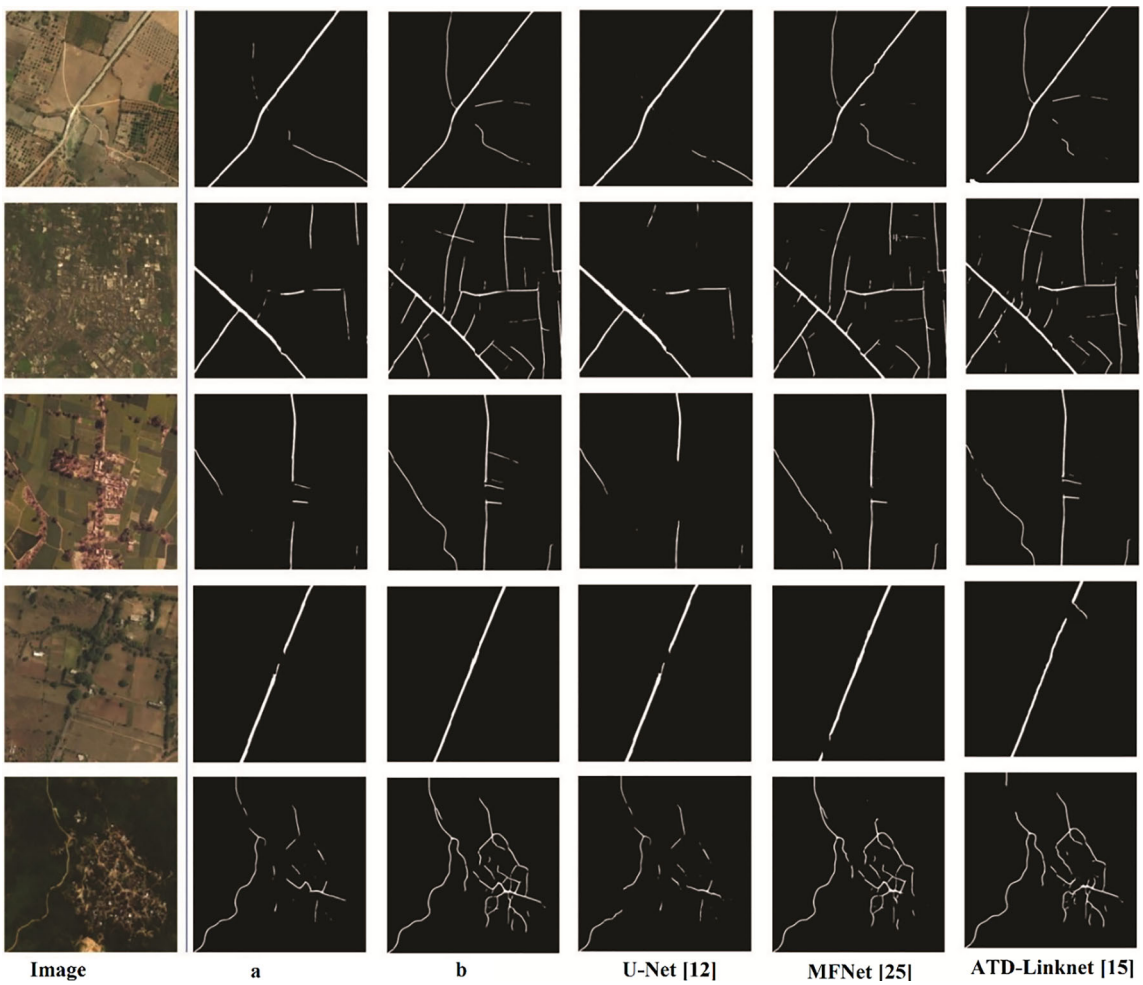


Fig. 3 Visual output of the proposed model in comparison with state-of-the art works on the DeepGlobe Road Extraction dataset. A represents the output of the U-Net–transformer with naive upsampling, and b represents the output of the U-Net–transformer with cascaded upsampling

Table 3 Influence of the number of skip connections on performances on the 2D ISPRS Potsdam dataset

Skip connection	Overall accuracy (%)	mIoU (%)
¼	89.34	82.88
¼ and ½	90.63	84.39
¼, ½, and 1/8	91.02	85.41

Table 4 Influence of the input resolution on performances on the 2D ISPRS Potsdam dataset

Input resolution	Overall accuracy (%)	mIoU (%)
512 × ×512	89.14	82.98
1024 × 1024	91.02	85.41

Table 5 Influence of the patch size on performances on the 2D ISPRS Potsdam dataset

Patch size	Overall accuracy (%)	mIoU (%)
8	91,02	85.41
16	90.98	85.29
32	90.92	85.11

presents the achieved results using large and light models. Considering the results, the light model was adopted due to its low computation complexity and low accuracy improvement.

The last ablation study was reserved for evaluating individual performances of the CNN model and transformers separately and the performances of the proposed hybrid model. Numerous experiments proved the superiority of the hybrid model compared to the separate CNN model or transformers (ViT). However, as both models

Table 6 Influence of the input resolution on performances on the 2D ISPRS Potsdam dataset

Model configuration	Overall accuracy (%)	mIoU (%)
Hidden layers: 12 MLP size: 3072 Number of heads: 12 Size of D: 768	90.74	84.38
Hidden layers: 24 MLP size: 6144 Number of heads: 24 Size of D: 1536	91.02	85.41

Table 7 Influence of the number of skip connections on performances on the 2D ISPRS Potsdam dataset

Model	Overall accuracy (%)	mIoU (%)
CNN	87.38	82.19
ViT	88.53	83.33
U-Net–transformer	91.02	85.41

Table 8 Achieved results of the proposed U-Net–transformer on the Synapse multi-organ segmentation dataset and the pascal VOC segmentation dataset

Dataset	Overall accuracy (%)	mIoU (%)
Synapse multi-organ	87.35	82.67
pascal VOC	79.12	68.84

present limitations in the segmentation of remote sensing images, the results of the CNN model and ViT do not outperform state-of-the-art works. Table 7 presents the results of the CNN, ViT and the proposed hybrid model.

4.3 Discussion

The proposed U-Net–transformer was composed of a decoder combining a CNN model and transformers and a decoder based on a cascaded upsampling technique. The proposed model was designed for the segmentation of remote sensing images. In addition, we proposed a novel hybrid model that relies on both models' capabilities to surpass each other's limitations. The proposed model was designed based on the U-Net style, and skip connections were considered to transmit low-level segmentation features to high-level features maps. The presented experiments have proved the robustness of the proposed model for the segmentation of remote sensing images. To benchmark, datasets were used for evaluation: the 2D ISPRS Potsdam dataset and the DeepGlobe Road Extraction dataset. As a result, the U-Net–transformer presented superior performances compared to state-of-the-art models. The proposed model as a medium computation complexity with 20.3 million parameters and a processing speed of 23 FPS on the Nvidia GTX 960 GPU. The processing speed can be boosted using better GPU with more processing units and memory.

4.4 Generalization to other segmentation tasks

To further demonstrate the performance of the proposed U-Net–transformer, it was evaluated on other segmentation tasks including the segmentation of natural and medical images. First, the proposed model was evaluated on the Synapse multi-organ segmentation dataset [33]. Also, the proposed model was evaluated on pascal VOC segmentation dataset [34]. Table 8 presents the achieved results on both datasets. Referring to the reported results, it was proved that the proposed model can be generalized other segmentation task such medical images and natural images without any modification to the structure. We believe that applying some modification may lead to better results for other segmentation task since the proposed U-Net–transformer was designed for very high-resolution images.

5 Conclusion

The importance of remote sensing images for various applications has been widely investigated recently. The main limitation is that those images are very complex and require considerable effort for analysis. Semantic segmentation of remote sensing images was proposed as a solution for fast and precise analyses. We proposed a semantic segmentation method for remote sensing images in this paper. The proposed method was based on a hybrid deep learning technique that combines a CNN model and transformers. Transformers are strong models based on attention mechanisms. CNN model can learn complex spatial features. The U-Net–transformer investigates the combination of CNN and transformers in the encoder stage and uses cascaded upsampling for the decoder stage. The model has a U-Net shape with skip connection at different resolution scales. The proposed model has achieved state-of-the-art performances on two benchmark datasets: the 2D ISPRS Potsdam dataset and the DeepGlobe Road Extraction dataset. The achieved results with 91.02% of accuracy and 85.41% of IoU on the 2D ISPRS Potsdam dataset in addition to 89.09% of accuracy and 83.87% of IoU on the

DeepGlobe Road Extraction dataset proved the efficacity of the proposed model and the efficiency of the combination of the CNN and the transformers. The main limitation of the proposed model is that it requires high-performance computation resources and the need for large memory. In future work, the model will be reconfigured to reduce the computation complexity, make it suitable for implementation on autonomous systems such as drones and meet real-time processing constraints.

Data availability Data will be made available on request.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

1. Brunn SD (2019) The international encyclopedia of geography: people, the earth, environment and technology. AAG rev Books 7(2):77–85
2. A. I. Godunov, Penza State University, S. T. Balanyan, P. S. Egorov, Air Force Academy named after Professor N. E. Zhukovsky and Yu. A. Gagarin, and Air Force Academy named after Professor N. E. Zhukovsky and Yu. A. Gagarin, 2021 "Image segmentation and object recognition based on convolutional neural network technology," Reliab. qual. complex syst., no. 3
3. P. Wang et al., 2018 "Understanding Convolution for Semantic Segmentation," In 2018 IEEE winter conference on applications of computer vision (WACV)
4. S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, 2018 "Path aggregation network for instance segmentation," In 2018 IEEE/CVF conference on computer vision and pattern recognition
5. M. N. Mullani and P. A. Dandavate. 2019 Semantic texton forests for image categorization and segmentation. Nternational j. adv. res. comput. commun. eng. 8(4): 259–262
6. Smith A (2010) Image segmentation scale parameter optimization and land cover classification using the Random Forest algorithm. J Spat Sci 55(1):69–79
7. Barthakur M, Sarma KK (2020) "Deep learning based semantic segmentation applied to satellite image", in Data Visualization and Knowledge Engineering. Springer International Publishing, Cham, pp 79–107
8. Ayachi R, Said Y, Atri M (2021) A convolutional neural network to perform object detection and identification in large-scale visual data. Big Data 9(1):41–52
9. Afif M, Ayachi R, Said Y, Atri M (2020) Deep learning based application for indoor scene recognition. Neural Process Lett 51(3):2827–2837
10. Ayachi R, Afif M, Said Y, Atri M (2020) Traffic signs detection for real-world application of an advanced driving assisting system using deep learning. Neural Process Lett 51(1):837–851
11. R. Ayachi, M. Afif, Y. Said, and A. B. Abdelaali, 2020 "Pedestrian detection for advanced driving assisting system: a transfer learning approach," In 2020 5th international conference on advanced technologies for signal and image processing (ATSIP)
12. Ronneberger O, Fischer P, Brox T (2015) "U-Net: Convolutional Networks for Biomedical Image Segmentation." Lecture Notes in Computer Science. Springer International Publishing, Cham, pp 234–241
13. X. Zhang, H. Yang, and E. F. Y. Young, "Attentional transfer is all you need: Technology-aware layout pattern generation," In 2021 58th ACM/IEEE design automation conference (DAC), 2021.
14. A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv [cs.CV], 2020.
15. Qi X, Li K, Liu P, Zhou X, Sun M (2020) Deep attention and multiscale networks for accurate remote sensing image segmentation. IEEE Access 8:146627–146639
16. L. Zhou, C. Zhang, and M. Wu, "D-LinkNet: LinkNet with pretrained encoder and dilated convolution for high-resolution satellite imagery road extraction," In 2018 IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW), 2018.
17. F. Yu and V. Koltun, "Multiscale context aggregation by dilated convolutions," arXiv [cs.CV], 2015.
18. Cui B, Chen X, Lu Y (2020) Semantic segmentation of remote sensing images using transfer learning and deep convolutional neural network with dense connection. IEEE Access 8:116744–116755
19. G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, 2017 "Densely connected convolutional networks," In 2017 IEEE conference on computer vision and pattern recognition (CVPR)
20. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. (2009) "ImageNet A large-scale hierarchical image database." In 2009 IEEE conference on computer vision and pattern recognition. Doi https://doi.org/10.1109/CVPR.2009.5206848
21. Gao L, Liu H, Yang M, Chen L, Wan Y, Xiao Z, Qian Y (2021) STransFuse: Fusing swin transformer and convolutional neural network for remote sensing image semantic segmentation. IEEE J Sel Top Appl Earth Obs Remote Sens 14:10990–11003
22. Zou Z, Shi T, Li W, Zhang Z, Shi Z (2020) Do game data generalize well for remote sensing image segmentation? Remote Sens 12(2):275
23. J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, 2017 "Unpaired image-to-image translation using cycle-consistent adversarial networks," In 2017 IEEE international conference on computer vision (ICCV)
24. K. He, X. Zhang, S. Ren, and J. Sun, 2016 "Deep residual learning for image recognition," In 2016 IEEE conference on computer vision and pattern recognition (CVPR)
25. Liu Y, Zhu Q, Cao F, Chen J, Gang Lu (2021) High-resolution remote sensing image segmentation framework based on attention mechanism and adaptive weighting. ISPRS Int J Geo Inf 10(4):241
26. Xu Z, Zhang W, Zhang T, Yang Z, Li J (2021) Efficient transformer for remote sensing image segmentation. Remote Sens 13(18):3585
27. Li A, Jiao L, Zhu H, Li L, Liu F (2021) Multitask semantic boundary awareness network for remote sensing image segmentation. IEEE Trans Geosci Remote Sens 60:1–14
28. F. Rottensteiner et al., 2012 "The isprs benchmark on urban object classification and 3d building reconstruction." ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences; I-3. 1(1): 293-298
29. M. Tan and Q. V. Le, 2021 "EfficientNetV2: Smaller models and faster training," arXiv [cs.CV]
30. D. Zhou et al., 2021 "DeepViT: Towards Deeper Vision Transformer," arXiv [cs.CV]
31. Liu, Ze, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021 "Swin transformer:

Hierarchical vision transformer using shifted windows." In Proceedings of the IEEE/CVF international conference on computer vision. 10012–10022
32. Demir, Ilke, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. 2018 "Deepglobe 2018: A challenge to parse the earth through satellite images." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 172–181
33. Zhang, Kaidong, and Dong Liu. (2023) "Customized segment anything model for medical image segmentation." arXiv preprint arXiv:2304.13785
34. Everingham M, Van Gool L, Williams CKI, Winn J, Zisserman A (2010) The pascal visual object classes (voc) challenge. Int j comput vis 88:303–338
35. Afaq Y, Manocha A (2021) Analysis on change detection techniques for remote sensing applications: a review. Eco Inform 63:101310
36. Bai T, Wang Le, Yin D, Sun K, Chen Y, Li W, Li D (2023) Deep learning for change detection in remote sensing: a review. Geospat Inform Sci 26(3):262–288
37. Zhang C, Wang L, Cheng S, Li Y (2022) SwinSUNet: Pure transformer network for remote sensing image change detection. IEEE Trans Geosci Remote Sens 60:1–13
38. Manocha A, Afaq Y (2023) Optical and SAR images-based image translation for change detection using generative adversarial network (GAN). Multimed Tools Appl 82(17):26289–26315