A review of AI transformers in image segmentation

Victor Doma[1,†], Erna Berbić [1,†], Ali Abd Almisreb[1,2], Mohammed A. Saleh[2,†] and Saule Amanzholova[3,†]

[1] *International University of Sarajevo, Faculty of Engineering and Natural Sciences, Hrasnička cesta 15, Ilidža, Bosnia and Herzegovina*

[2] *International Information Technology University, 34/1 Manas St., Almaty, 050000, Kazakhstan*

[3] *Astana IT University, Astana, Kazakhstan*

Abstract

In digital image processing and analysis, image segmentation is a widely used technique that divides a picture into multiple different sections or areas, usually depending on the properties of the image's pixels. Segmenting an image might include grouping pixel sections according to color or form similarity, or it could lead to dividing the foreground from the background. A transformer model is a type of neural network that tracks connections in sequential data, such as the words in a phrase, to determine context and meaning. In recent years, transformers have significantly outperformed earlier convolutional or recurrent methods in a variety of visual processing applications. This paper provides a thorough overview of AI transformers in image segmentation. Firstly, providing a literature review of the topic, then diving deep into the various transformer methodologies used in image segmentation. Lastly, we compile and discuss the reviewed methods, identify challenges and purpose directions for future research.

Keywords

AI Transformers, Image Segmentation, Deep Learning, Computer Vision

1. Introduction

Image segmentation is the very first step for image analysis and pattern recognition. Image analysis and pattern recognition rely heavily on this difficult challenge, which ultimately affects the quality of the analysis results. Image segmentation divides an image into distinct sections that are homogenous, but not when combined with adjacent parts [1]. Deep neural networks, including Convolution Neural Networks (CNNs) and Fully Convolutional Networks (FCNs) have significantly improved segmentation achievements throughout the last decade [2]. CNN-based segmentation outperforms classical techniques in terms of generalization [3]. As a result, CNN architectures are widely used in segmentation studies because to their high performance [4], [5]. The rising popularity of Natural Language Processing (NLP) has also led to the debut of transformer as a replacement for recurrent neural networks [6], [7].

Modern image segmentation approaches rely on transformer architecture, where transformer-based techniques outperform CNN in terms of pipeline simplicity and performance. Because of their rapid growth, recent studies have been made [8], [9], mostly focusing on generic transformer design and its application to particular vision problems [10], [11]. There have also been prior surveys on deep-learning-based segmentation.

The contributions which we provide throughout this paper are:

DTESI 2024: 9th International Conference on Digital Technologies in Education, Science and Industry, October 16–17, 2024, Almaty, Kazakhstan

* Corresponding author.

† These authors contributed equally.

vdoma@student.ius.edu.ba (V. Doma); eberbic@student.ius.edu.ba (E. Berbić); alimes96@yahoo.com (A. Almisreb); s.amanzholova@astanait.edu.kz (S. Amanzholova)

0000-0001-7581-5747 (A. Almisreb); 0000-0002-6779-9393 (S. Amanzholova)

1. **Comprehensive Analysis**: we bring an in-depth analysis of the impact AI transformers have made in the field of image segmentation and how they, in a way, revolutionized the field.
2. **Comparative Evaluation**: even though traditional approaches, such as CNN and FCN, have also left a positive mark on image segmentation, this work focuses on highlighting what kind of contributions and gains AI transformers have made.
3. **Encouraging Future Research**: we hope that by offering insights into the untapped potential and future prospects of transformer-based segmentation methodologies, we will be able encourage and inspire more research and innovation in this rapidly developing sector.

The study is organized as follows. In the first section, a compilation of various research papers has been made with the aim of providing a thorough analysis of each one of them to help us further understand how the architecture of these transformers has helped revolutionize the field of image segmentation. Furthermore, the purpose is to also point out, when applied, how transformers have the ability to improve the overall accuracy and efficiency. The second section offers a very in-depth analysis of various numerous types of AI transformers, tapping into the architecture and their benefits. Further discussing, through a comparative analysis, as to in which applications or more specifically in which segmentation tasks they can be implemented and taking into account the benefits they bring, but also the limitations they possess.

2. Literature review

Image segmentation has entered a new era with the appearance and utilization of AI transformers. Although traditional methodologies still play a significant role in image segmentation, with the integration of transformers, new horizons for advancing segmentation methodologies are opened for us to explore. In this literature review, we explore what kind of an impact have AI transformers made in this field. Throughout this review, the objective is to provide valuable insight into the progressive world of image segmentation and with it inspire additional future research to be made. In his study, Minh Tran introduces Amodal Instance Segmentation (AIS) and the challenges that go with it, since it involves predicting both visible and, so to say, hidden parts of objects within images. The study reveals that current AIS techniques, based on bidirectional approaches, can lead to confusion between the visible and amodal features. The author proposes ShapeFormer as a solution to this problem, which is decoupled Transformer-based model with a visible-to-amodal transition. This gives us a clear relationship between output segmentations as well as eliminating the requirement for amodal-to-visible transitions [12].

The study in[13] explored the use of a pure Vision Transformer for the image segmentation of remote sensing images. The experimental results proof that general-purpose transformer models are highly competitive even against state-of-the-art transformers. Zhaoyang Ma discusses enhancing rock image segmentation in digital rock physics, whilst also touching upon the drawbacks of traditional segmentation approaches. Their study demonstrates an advanced generative AI model, also known as the diffusion model, which overcomes many limitations. The model produced a large number of CT/SEM and binary segmentation pairings from a modest starting sample. They also incorporated into their study a performance comparison analysis between U-Net, Attention-U-Net and TransUNet, where the diffusion model has proved to be an effective data augmentation technique, improving the generalization and robustness of deep learning models [14].

In his study Hazrat Ali focuses on how vision transformer-based methods are climbing the ladder in the medical AI area. More specifically, lung cancer imaging. A synthesis of the collected literature was made in order to perform adequate lung cancer type classification. In other words, distinguish benign and malignant pulmonary nodules, discussing the positive and effective use of transformer-based approaches in such cases [15, 16]. In his paper, Zhou Deng introduces transformer-based generative adversarial network for real fundus images restoration, in hopes of making a new clinical benchmark. The study introduces RFormer, which is a transformer-based generative adversarial

network, demonstrating its superior ability to restore fundus images and downstream tasks like vessel segmentation and optic disc/cup recognition, emphasizing its potential for clinical analysis and many other applications as well [17].

To conclude the literature review, the purpose of it was to establish some context on the topic discussed in this paper, which helps the reader get a better grasp of the background and significance of the study. Another benefit gained from these reviews is discovering gaps, inconsistencies, or unresolved questions in the research, which helps us define the scope and the direction of our study.

3. Background

3.1. Convolutional neural networks for image segmentation

A convolutional neural network (CNN) is an instance of artificial neural network that is mostly used for image recognition and processing since it has the ability to detect patterns in imagery. The architecture of the model is made out of convolutional layers, pooling layers and activation functions. The input layer is the layer inside of CNN which holds the image's pixel value. The convolutional layer calculates the scalar product of neurons' weights and the input volume to determine their output. The ReLu applies an activation function, such as sigmoid, to the output of the preceding layer. The pooling layer will execute downsampling along the spatial dimension of the input, lowering the number of parameters inside that activation. The fully-connected layers will generate class scores from the activations, in order to be used as classification. ReLu might be applied between these layers as to improve overall performance [18].

Popular CNN architectures for image segmentation include U-Net and Fully Convolutional Networks (FCNs). U-Net is able to skip connections which allows for more exact localization in segmentation tasks [19] and FCNs use fully convolutional layers to build segmentation maps directly from input images, which allows for complete end-to-end training for segmentation tasks [20].

3.2. Transformers and self-attention mechanisms

Transformers rely on attention to establish global relationships between input and output, rather than repetition. They enable high parallelization and achieve exceptional translation quality after just 12 hours of training on eight P100 GPUs [6]. Transformers use self-attention layers, similar to Non-Local Neural Networks [21], to update a sequence by aggregating input from all elements. Attention-based models excel at extended sequences because of their global calculations and flawless memory, outperforming RNNs [22]. Transformers are increasingly replacing RNNs in natural language processing, voice processing, and computer vision [23]. Following the success of Transformer in NLP, several works introduced self-attention to CNN, showing that self-attention and CNN can be integrated or replaced with each other [24], [25]. Later on, researchers have started exploring the possibility of removing the convolutional layer as the core block, which then proved to be effective. Experiments demonstrate that self-attention may effectively replace the convolutional layer and improve the performance of image segmentation [26].

3.3. Advantages and limitations of transformers

Transformers provide many advantages when utilized in computer vision, more specifically image segmentation. For starters, they are highly parallelizable, which means they can analyze several sections of a sequence concurrently, considerably speeding up training and inference. Furthermore, transformers can detect long-term relationships, allowing them to better grasp the broader context and produce more cohesive content. They are also more adaptable and scalable, making them simpler to apply to many activities and domains [23], [27], [28].

As for the limitations, one of the main disadvantages is the high computational requirement. Transformer-based models demand significant computer resources and training time due to their size

and complexity. Since transformers are highly sensitive to the quality and quantity of training data, if the training data is restricted or biased, the model's performance may suffer [6], [29], [30].

4. Methodology

This study reviews numerous research articles on AI transformers in image segmentation, from their origins, architecture, applications and recent advancements. We thoroughly analyze the pertinent research articles published on this subject matter, evaluating the datasets, preprocessing, materials, and methods used and the acquired results. We compare and discuss the successes and limitations of each work, highlighting their main objectives and their impact on image segmentation over the years. These transformers employ different strategies to perform image segmentation and based on their methods of adaptability, we group the transformers into distinct categories, covering methods such as modifying the fundamental architecture of models expressly for segmentation tasks and adding segmentation heads to pre-trained models. Using common image segmentation benchmarks, our study evaluates how well these different transformers perform. This makes it possible to compare how effective they are. Figure 1 shows the different categories of transformers used in image segmentation.

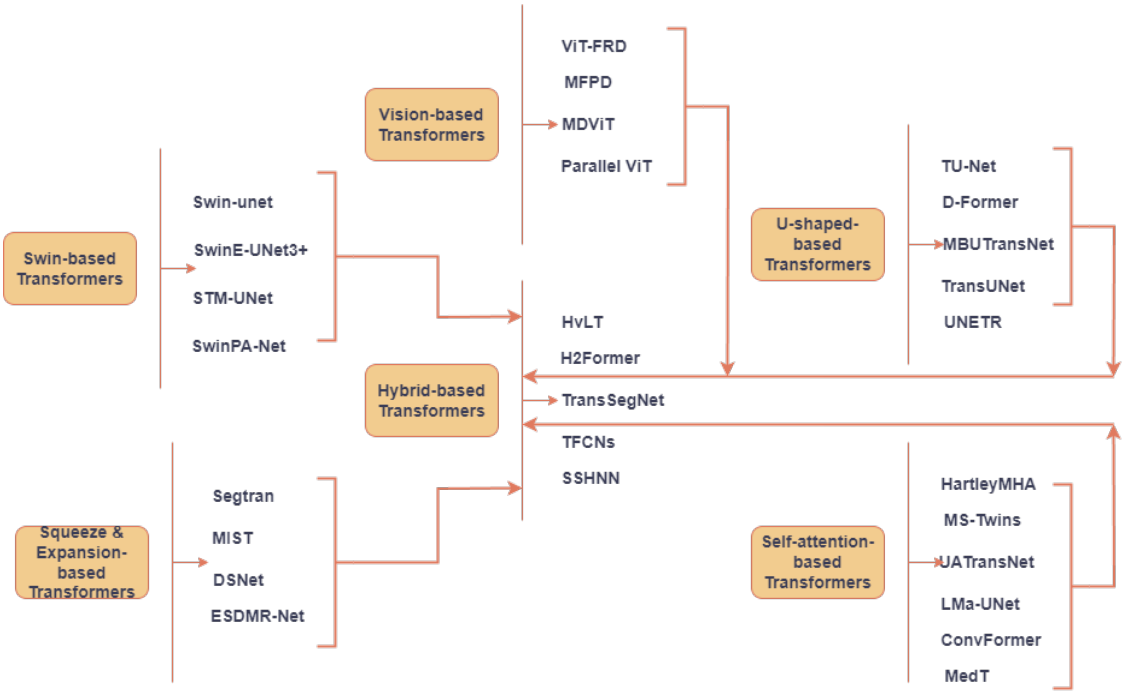

Figure 1: Our graphical representation of the different categories of transformers used in image segmentation tasks.

4.1. Vision-based transformers

To create a Visual Transformer with Feature Recombination and Feature Distillation (ViT-FRD) for MRI image segmentation, architectures of the vision transformer (ViT) and CNN are joined together [31]. Through optimizing distillation losses [32], ViT learns from CNN in this transformer model. ViT-FRD achieves great results on two cardiac MRI image datasets, surpassing baseline models. To enable enhanced segmentation in the automatic analysis of remote sensing image data, a multiscale feature pyramid decoder (MFPD) is presented in [33]. Convolution and single-scale feature maps [34] struggle to segment large amounts of remote sensing images are often difficult to segment because of varying patterns. The vision transformer decoder uses a 2-D-to-3-D transform approach to extract rich multiscale feature maps and together with a dimension attention module (DAM) [35] binds the image features. Hence the model achieves high mean intersection over union (mIoU) values on the

Gaofen2-CZ dataset [36] and GID-5 dataset [37]. An innovative multi-domain ViT (MDViT) technique for medical image segmentation is presented in [38]. Because vision transformers are often trained from one data source, they fail to identify essential information contained in other datasets, resulting in a negative knowledge transfer (NKT). MDViT consists of domain adapters used to reduce the need for large amounts of data and counter NKT by using knowledge from domains (numerous small datasets), so expectedly MDViT outperformed many state-of-the-art algorithms.

To overcome the limitations of a typical CNN and sequential Transformer model in medical image segmentation, PTransUNet and C-PTransUNet models were proposed in [39]. The C-PT unit improves ViT by replacing its sequential architecture with a parallel one, enhancing its feature extraction capabilities. These models achieve higher model accuracy compared with the baseline model on the Synapse dataset. To enable higher image segmentation accuracy for detecting very small or many targets in overlapping sections, a vision transformer with unified-perceptual-parsing network (ViT-UperNet) was presented [40]. The model uses a ViT fixed with a self-attention mechanism that extracts image features in a hierarchy approach and implements a unified-perceptual-parsing network [41] for feature fusion and image segmentation.

4.2. Swin-based transformers

A novel Unet-like Transformer named Swin-Unet, for medical image segmentation is presented in [42]. To achieve feature learning the model passes the tokenized image patches [43] into a hierarchical shifted windows Swin Transformer's U-shaped Encoder-Decoder. And a symmetric Swin Transformer-based decoder with a patch expanding layer in the Swin-based decoder up-samples image features restoring quality feature maps resolution [44]. To improve contour details in tumor segmentation SwinE-UNet3+ model is proposed in [45]. Using the two consecutive Swin Transformer blocks, each with its own SwinE-UNet3+ encoder layer, enables the extraction of long-range image features using patch merging. The decoder employs the Conv2DTranspose feature up-sampling [46] and convolution operation to combine the decoder and encoder information. Hence the model is applied on the TipDM Cup rectal cancer dataset and the melanoma dermoscopic image ISIC-2017 dataset [47], achieving better Dice coefficient [48], IOU value and Precision values than UNet, UNet++ [49] and UNet3+ [50].

A highly efficient U-shaped architecture based on Swin Transformer and multiscale MLP (STM-UNet) is presented in [51]. To ensure rich global features and improve long-range dependencies, the Swin Transformer is added to STM-UNet skip connections. To improve segmentation the authors designed a parallel convolution block in axial-shifted machine learning perceptron (PCAS-MLP) module and placed it in the proposed model, which achieved superior IoU and Dice results, compared to other state-of-the-art techniques. Big differences between different types of lesions and similar colors and shapes between lesions and tissues affect segmentation accuracy. To overcome this challenge, the authors present, Swin Pyramid Aggregation network (SwinPA-Net) [52]. The network is a combination Swin Transformer and the dense multiplicative connection (DMC) module and local pyramid attention (LPA) modules, used for aggregating the multiscale context information of images. The network is evaluated on the polyp segmentation and skin lesion segmentation datasets, achieving greater results than some existing state-of-the-art (SOTA) methods.

4.3. U-shaped-based transformers

To perform 3D medical image segmentation a novel approach named UNEt TRansformer (UNETR) is presented in [53]. It employs a transformer encoder to extract multi-scale global information and a U-shaped encoder and decoder. The model is validated on the Multi Atlas Labeling Beyond the Cranial Vault dataset [54] and the Medical Segmentation Decathlon dataset. For 3D medical image segmentation, a D-Former model based on a Dilated Transformer and a U-shaped encoder–decoder is proposed in [55]. The novel Dilated Transformer has a dilated self-attention module for enlarging

image patch receptive and lessening computational costs. The model is validated on the Synapse and ACDC datasets and achieves better results than CNN-based and Transformer-based models.

A novel method for medical cell segmentation and abdominal organs segmentation using a multibranch U-shaped structure fusion transformer network (MBUTransNet) is introduced in [56]. The model consists of the coordinate attention transformer [57], designed for extracting long-term dependency information and small U-net blocks and a multiscale feature fusion block to combine multi-layer feature maps. The results demonstrate that on the MoNuSeg [58] and Synapse multiorgan segmentation datasets, MBUTransNet acquires a 0.076 and 0.1269 DICE improvement, respectively. To create a better alternative for medical image segmentation, TransUNet, a model that consists of Transformers and U-Net is proposed in [59]. To obtain global context information from feature maps, the Transformer encodes tokenized image patches. The decoder upsamples the features and combines them with the rich CNN feature maps to improve precise localization. A novel technique, TU-Net based on transformers, to overcome the limitations of U-Net is proposed in [60]. TU-Net improves the extraction of global context information and decreases the model's computational complexity by using patch embedding. To combine the image features they created a cross attention-skip module [61]. TU-Net is performed on the Synapse dataset to segment eight abdominal organs. The results show that TU-Net outperforms ViT, V-Net, U-Net and Swin-Unet.

4.4. Self-attention-based transformers

To perform 3D image segmentation, the HartleyMHA model with efficient self-attention mechanism was proposed in [62]. Fourier Neural Operator (FNO) [63], a deep learning architecture for integrating mappings between functions in partial differential equations is modified using the Hartley transform [64] to enhance model performance and decrease model size. The model is validated on the BraTS'19 dataset [65] outperforming other models. MS-Twins (Multi-Scale Twins), a model that combines self-attention and convolution to perform medical image segmentation is introduced in [66]. The model joins multi-scale features for richer image information. The model is validated on the Synapse and ACDC datasets surpassing SwinUNet by 8%. To perform osteosarcoma MRI image segmentation a lightweight image segmentation network, UATransNet is proposed in [67]. The network employs a multilevel guided self-aware attention module (MGAM) and a U-Net encoder-decoder. The network's transformer self-attention component (TSAC) and global context aggregation component (GCAC) integrate the local features and global dependencies. To enhance feature extraction the authors, apply dense residual learning to the convolution module. UATransNet very high IOU and DSC scores, 0.922 ± 0.03 and 0.921 ± 0.04, respectively. To achieve efficient 2D and 3D medical image segmentation, a Large Window-based Mamba U-shaped Network (LMa-UNet) was proposed in [68]. The network's large windows and an innovative hierarchical and bidirectional Mamba module enhance spatial modeling and is efficient in global modeling. Experimental results highlight the network's high efficiency.

An innovative CNN-based Transformer (ConvFormer) technique for medical image segmentation is proposed in [69]. The model uses 2D convolution and max-pooling for retaining position information and decreasing the feature size. The model employs a CNN-style self-attention (CSA) which makes self-attention matrices to establish long-range dependency, then convolutional feed-forward network (CFFN) feature refinement. A novel pyramidal network architecture of multi-scale attention and CNN feature extraction dubbed Pyramid Medical Transformer (PMTrans) is presented in [70]. The model uses multi-resolution images and an adaptive partitioning approach was implemented to efficiently retain position information relations and to interact with diverse receptive fields. PMTrans was validated on and acquired very good results on the gland segmentation, MoNuSeg, and HECKTOR datasets [71]. A novel gated axial-attention model named Medical Transformer (MedT), for medical image segmentation is introduced in [72]. The model has a control mechanism added to the its self-attention module and for enhanced model training a local-global training technique (LoGo) [73] is used. Specifically, we operate on the whole image and patches to

learn Global and local features are learned by processing the whole image and image patches. MedT achieves superior experimental to those the convolutional and other transformer-based models.

4.5. Squeeze and expansion-based transformers

To perform efficient medical image segmentation, the authors presented Segtran, a transformer with unlimited effective receptive fields [74]. The core principle of the model based on a Squeeze-and-Expansion transformer. Here, a squeezed attention block [75] regularizes the self-attention of transformers, and an expansion block learns different image representations. The authors also present a positional encoder that enables a continuity inductive bias for images. The model is validated on the REFUGE'20 [76] challenge, polyp segmentation and BraTS'19 challenge datasets. A Medical Image Segmentation Transformer (MIST) consisting of an innovative convolutional attention mixing (CAM) [77] is presented in [78]. MIST consists of a pre-trained multi-axis vision transformer (MaxViT) encoder and a CAM decoder that joins multi-head self-attention, spatial attention, and squeeze and excitation attention modules for long-range dependencies extraction. Deep and shallow convolutions are employed for enhanced feature extraction. The model produces better results than some state-of-the-art models specifically designed for medical image segmentation.

An efficient convolutional neural network (CNN) and transformer, known as Dynamic Squeeze Network (DSNet), is proposed for real-time weld seam segmentation in [79]. The model comprises of an efficient encoder for different features and a novel plug-and-play lightweight attention module that creates more effective attention weights by using linear priors. DSNet significantly decreases the number of parameters, computational complexity while increasing inference speed, compared to TransUNet. An expand-squeeze dual multiscale residual network (ESDMR-Net) for medical image segmentation is proposed in [80]. The model performs a dual encoder–decoder information flow. The expansion operation extracts the rich multi-scale features for improve segmentation. The Expand-Squeeze (ES) module enhances segmentation accuracy by focusing on the under-represented classes. The dual multiscale residual (DMR) [81] modules enable multi-scale information flow using skip connections. ESDMR-Net was validated on seven datasets and achieved high f1 scores of 0.8287%, 0.8211%, 0.9034%, 0.9451%, 0.9543%, 0.9840%, and 0.8424% on the DRIVE, CHASE, ISIC2017, ISIC2016, CVC-ClinicDB, MC and MoNuSeg datasets, respectively.

4.6. Hybrid-based transformers

A hybrid ladder transformer (HyLT) for medical image segmentation is introduced in [82]. To encode long-range dependencies which are fused by bi-directional cross attention module [83], the model employs global attention heads conjoined with a CNN. The model is validated on two medical image datasets. To implement efficient medical image segmentation, a novel hierarchical hybrid vision Transformer (H2Former) is presented in [84]. The model constitutes of Transformers and multi-scale channel attention. Experimental results show that the model is highly efficient even with limited medical data, even surpassing TransUNet by an IoU of 2.29% on the KVASIR-SEG dataset [85]. An innovative hybrid network, TranSegNet for retina segmentation is proposed in [86]. The network consists of lightweight ViT with a multi-head convolutional attention and a U-shaped Transformer based backbone for global feature extraction, accurately localizing retinal layers and lesion tissues. This hybrid CNN-ViT model achieves high efficiency and accuracy in the segmentation of retinal layers and accumulated fluid and outperformed FCN, SegNet, Unet and TransUnet.

An innovative hybrid approach for medical image segmentation, Transformers for Fully Convolutional denseNets (TFCNs) is presented in [87]. The model consists of ResLinear-Transformer (RL-Transformer) and convolutional linear attention block (CLAB) to FC-DenseNet. TFCNs use the latent information from CT images to perform feature extraction. The model is validated on the Synapse dataset and acquires a high Dice score of 83.72%. A semi-supervised hybrid NAS network

named SSHNN is introduced for efficient medical image segmentation [88]. The network uses the convolution operation in layer-wise feature fusion to enhance NAS's encoding capabilities. Transformers are used to retain global context and a U-shaped decoder links global context with the local features. Experimental results on the CAMUS echocardiography dataset [89] demonstrates how SSHNN outperforms other state-of-the-art methods.

5. Transformer model selection based on image type

It is important to emphasize that choosing the right transformer architecture is contingent on the nuances found in the particular image segmentation task. Understanding the task requirements such as type of images and the resolution and complexity of the images is the prerequisite: for instance, on the one hand, some tasks demand high precision and structural preservation, and on the other, some emphasize boundary detection. High-resolution images or images with complicated textures could benefit from the transformer models that can process very fine detailed features.

Medical images almost always contain fine granularity features (i.e., cell boundaries, tissues). Swin-based transformers, Squeeze-and-Expansion-based transformers, Self- Attention based Transformers and U shape-based transformers are preferred because of their ability to capture local and global structures sufficiently, particularly preserving the hierarchical detail. For aerial images that capture both large-scale structures like buildings, landscapes and roads, Vision-based transformers, Self-Attention based Transformers and Swin-based Transformers are often used. Images that require the segmentation of natural scenes like animals and people, Self-Attention based Transformers and Vision Transformers are generally used, where long-range dependencies and various object sizes are important. For images with multiple objects and scales, such as indoor room layouts or outdoor street views, almost always benefit from Swin-based transformers and Squeeze-and-Expansion-based transformers. Hybrid-based transformer models are usually a combination of CNN and a transformer or at least two transformers. They are used for tasks requiring both local feature extraction and global context (e.g., medical images). They generally perform well for most image segmentation tasks but selecting the appropriate combination with regards to the task is essential.

6. Discussion

In computer vision research, the use of transformers architectures first intended for NLP for image segmentation has gained a lot of attention. Transformers have special benefits that are transforming the way we approach image segmentation challenges. Transformer-based CNNs are excellent at capturing the associations between far-off image regions, enabling precise segmentation of complicated objects with intricate geometries or interactions with the background, this is essential. Since transformers can interpret an image in its entirety at once, they are able to comprehend the image's overall context and the relationships between its many components. Accurate image segmentation is improved by this comprehensive understanding, particularly for tasks like panoptic segmentation that call on both localization and object classification. By combining and adapting the transformer architecture with different modules, researchers can create models that are specifically tailored for distinct segmentation tasks or types of data.

In comparison to CNNs, standard transformer designs may be more computationally costly. Real-time apps or their deployment on devices with limited resources may encounter difficulties as a result. The development of transformer-based segmentation models that are both lightweight `and efficient is the main goal of research efforts. A significant amount of labelled data is frequently needed to train transformer models effectively. This may provide a problem in certain domains where there may not be as much labelled data. To get around this restriction, methods like data augmentation and transfer learning are being investigated. Transformers are not necessarily better than CNNs in segmentation tasks, even though they have many advantages. Hybrid techniques,

which combine the advantages of transformers and CNNs, using transformers for global context modeling and CNNs for effective local feature extraction are highly efficient.

7. Conclusion

To conclude, this paper delves deep into the impact of the application of AI transformers in image segmentation, covering the obtained literature and a wide range of transformer models and the impact they have. Each model which was mentioned (more specifically, studied) throughout this paper offers unique methodologies and innovations for tackling image segmentation tasks, highlighting the strengths and weaknesses as well as shedding light on their effectiveness in various scenarios by evaluating the quantitative and qualitative results of the various AI transformers in image segmentation. With this review, we wanted to showcase the effectiveness, efficiency and extendibility transformers offer to image segmentation when applied. Overall, there is a bright future for transformer-based models, potentially shifting the landscape of computer vision research and application as it continues to evolve.

8. Future work

Researchers are currently investigating strategies to maintain the strengths of transformers while increasing their efficiency. This covers methods such as model pruning and sparse attention strategies. Gaining confidence in transformer models' applications requires an understanding of how they make segmentation decisions. Developing transformer-based models specifically designed for various segmentation tasks, like medical image or video segmentation, is a promising avenue for further advancement. Transparency and dependability can be improved by conducting research on interpretable transformers (Explainable AI) for picture segmentation.

Acknowledgements

We would like to show great appreciation for Professor Ali Abd Almisreb, for all his guidance and support throughout the whole process of this research. We would also like to extend our gratitude to the International University of Sarajevo for equipping us with all the necessary facilities and resources to perform this study.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] H. D. Cheng, X. H. Jiang, Y. Sun, and J. Wang, “Color image segmentation: advances and prospects,” Pattern Recognition Society, 2001.

[2] X. Li et al., “Transformer-Based Visual Segmentation: A Survey,” IEEE Trans Pattern Anal Mach Intell, Apr. 2023, [Online]. Available: http://arxiv.org/abs/2304.09854.

[3] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” Published as a conference paper at ICLR, Sep. 2014, [Online]. Available: http://arxiv.org/abs/1409.1556.

[4] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” Dec. 2015, [Online]. Available: http://arxiv.org/abs/1512.03385.

[5] Dolz, Jose, et al. “HyperDense-Net: a hyper-densely connected CNN for multi-modal image segmentation.” IEEE transactions on medical imaging 38.5 (2018): 1116-1126.

[6] A. Vaswani et al., “Attention Is All You Need,” 31st Conference on Neural Information Processing Systems, Jun. 2017, [Online]. Available: http://arxiv.org/abs/1706.03762.

[7] S. Hochreiter and J. Schmidhuber, “LONG SHORT-TERM MEMORY,” Neural Comput, 1997.
[8] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, “Transformers in Vision: A Survey,” ACM Comput Surv, Jan. 2022, doi: 10.1145/3505244.
[9] K. Han et al., “A Survey on Visual Transformer,” TPAMI, Dec. 2023, doi: 10.1109/TPAMI.2022.3152247.
[10] J. Lahoud et al., “3D Vision with Transformers: A Survey,” 2022. [Online]. Available: https://github.com/lahoud/3d-vision-transformers.
[11] B. Ni et al., “Expanding Language-Image Pretrained Models for General Video Recognition,” 2022.
[12] M. Tran, W. Bounsavy, K. Vo, A. Nguyen, T. Nguyen, and N. Le, “ShapeFormer: Shape Prior Visible-to-Amodal Transformer-based Amodal Instance Segmentation,” Mar. 2024, [Online]. Available: http://arxiv.org/abs/2403.11376.
[13] Gonçalves, Miguel, Bruno Martins, and Jacinto Estima. "A Detailed Analysis on the Use of General-purpose Vision Transformers for Remote Sensing Image Segmentation." Proceedings of the 6th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery. 2023.
[14] Z. Ma, X. He, H. Kwak, J. Gao, S. Sun, and B. Yan, “Enhancing Rock Image Segmentation in Digital Rock Physics: A Fusion of Generative AI and State-of-the-Art Neural Networks,” 2023.
[15] H. Ali, F. Mohsen, and Z. Shah, “Improving diagnosis and prognosis of lung cancer using vision transformers: A scoping review,” BMC Medical Imaging journal, 2023.
[16] Zhou, Hong-Yu, et al. "A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics." Nature biomedical engineering 7.6 (2023): 743-755.
[17] Z. Deng et al., “RFormer: Transformer-based Generative Adversarial Network for Real Fundus Image Restoration on A New Clinical Benchmark,” IEEE J Biomed Health Inform, Jan. 2022, [Online]. Available: http://arxiv.org/abs/2201.00466.
[18] K. O’Shea and R. Nash, “An Introduction to Convolutional Neural Networks,” Nov. 2015, [Online]. Available: http://arxiv.org/abs/1511.08458.
[19] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” May 2015, [Online]. Available: http://arxiv.org/abs/1505.04597.
[20] E. Shelhamer, J. Long, and T. Darrell, “Fully Convolutional Networks for Semantic Segmentation,” May 2016, [Online]. Available: http://arxiv.org/abs/1605.06211.
[21] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local Neural Networks,” Nov. 2017, [Online]. Available: http://arxiv.org/abs/1711.07971.
[22] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-End Object Detection with Transformers,” May 2020, [Online]. Available: http://arxiv.org/abs/2005.12872.
[23] N. Parmar et al., “Image Transformer,” Feb. 2018, [Online]. Available: http://arxiv.org/abs/1802.05751.
[24] R.-Y. Ju, T.-Y. Lin, J.-S. Chiang, J.-H. Jian, Y.-S. Lin, and L.-R.-Y. Huang, “Aggregated Pyramid Vision Transformer: Split-transform-merge Strategy for Image Recognition without Convolutions,” 2021.
[25] D. Mahajan et al., “Exploring the Limits of Weakly Supervised Pretraining,” 2018.
[26] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, “Stand-Alone Self-Attention in Vision Models,” Jun. 2019, [Online]. Available: http://arxiv.org/abs/1906.05909.
[27] A. Dosovitskiy et al., “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” arXiv preprint arXiv:2010.11929, Oct. 2020, [Online]. Available: http://arxiv.org/abs/2010.11929.
[28] Y. Xu, Q. Zhang, J. Zhang, and D. Tao, “ViTAE: Vision Transformer Advanced by Exploring Intrinsic Inductive Bias”.

[29] P. Chrabaszcz, I. Loshchilov, and F. Hutter, “A DOWNSAMPLED VARIANT OF IMAGENET AS AN AL-TERNATIVE TO THE CIFAR DATASETS”, Accessed: Apr. 04, 2024. [Online]. Available: https://github.com/PatrykChrabaszcz/Imagenet32_Scripts.
[30] I. Bello, B. Zoph, A. Vaswani, J. Shlens Quoc, and V. Le Google Brain, “Attention Augmented Convolutional Networks”.
[31] Fan, Chunyu, et al. "ViT-FRD: A vision transformer model for cardiac MRI image segmentation based on feature recombination distillation." IEEE Access (2023).
[32] Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. "Distilling the knowledge in a neural network." arXiv preprint arXiv:1503.02531 (2015).
[33] Wang, Wei, et al. "A ViT-based multiscale feature fusion approach for remote sensing image segmentation." IEEE Geoscience and Remote Sensing Letters 19 (2022): 1-5.
[34] Shi, Lei, Xiang Xu, and Ioannis A. Kakadiaris. "SSFD: A face detector using a single-scale feature map." 2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS). IEEE, 2018.
[35] Cai, Silin, et al. "FDAM: full-dimension attention module for deep convolutional neural networks." International Journal of Multimedia Information Retrieval 11.4 (2022): 599-610.
[36] Tong, Xin-Yi, et al. "Land-cover classification with high-resolution remote sensing images using transferable deep models." Remote Sensing of Environment 237 (2020): 111322.
[37] Yang, Kunping & Tong, Xin-Yi & Xia, Gui-Song & Shen, Weiming & Zhang, Liangpei. (2022). Hidden Path Selection Network for Semantic Segmentation of Remote Sensing Images. IEEE Transactions on Geoscience and Remote Sensing. 1-1. 10.1109/TGRS.2022.3197334.
[38] Du, Siyi, et al. "Mdvit: Multi-domain vision transformer for small medical image segmentation datasets." International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer Nature Switzerland, 2023.
[39] Wang, Dong, et al. "Cross-Parallel Transformer: Parallel ViT for Medical Image Segmentation." Sensors 23.23 (2023): 9488.
[40] Ruiping, Yang, et al. "ViT-UperNet: a hybrid vision transformer with unified-perceptual-parsing network for medical image segmentation." Complex & Intelligent Systems (2024): 1-13
[41] Xiao, Tete, et al. "Unified perceptual parsing for scene understanding." Proceedings of the European conference on computer vision (ECCV). 2018.
[42] Cao, Hu, et al. "Swin-unet: Unet-like pure transformer for medical image segmentation." European conference on computer vision. Cham: Springer Nature Switzerland, 2022.
[43] Wu, Bichen, et al. "Visual transformers: Token-based image representation and processing for computer vision." arXiv preprint arXiv:2006.03677 (2020).
[44] Hamwood, Jared, et al. "Effect of patch size and network architecture on a convolutional neural network approach for automatic segmentation of OCT retinal layers." Biomedical optics express 9.7 (2018): 3049-3066.
[45] Zou, Ping, and Jian-Sheng Wu. "SwinE-UNet3+: swin transformer encoder network for medical image segmentation." Progress in Artificial Intelligence 12.1 (2023): 99-105.
[46] Tang, Zeming, et al. "DenseNet with Up-Sampling block for recognizing texts in images." Neural Computing and Applications 32 (2020): 7553-7561.
[47] Berseth, Matt. (2017). ISIC 2017 - Skin Lesion Analysis Towards Melanoma Detection.
[48] Baudin, P-Y & Azzabou, Noura & Carlier, Pierre & Paragios, Nikos. (2012). Automatic skeletal muscle segmentation through random walks and graph-based seed placement. Proceedings - International Symposium on Biomedical Imaging. 1036- 1039. 10.1109/ISBI.2012.6235735.
[49] Zhou, Zongwei, et al. "Unet++: A nested u-net architecture for medical image segmentation." Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4. Springer International Publishing, 2018.

[50] Huang, Huimin, et al. "Unet 3+: A full-scale connected unet for medical image segmentation." ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2020.
[51] Shi, Lei, et al. "STM-UNet: An Efficient U-shaped Architecture Based on Swin Transformer and Multiscale MLP for Medical Image Segmentation." GLOBECOM 2023-2023 IEEE Global Communications Conference. IEEE, 2023.
[52] Du, Hao, et al. "SwinPA-Net: Swin transformer-based multiscale feature pyramid aggregation network for medical image segmentation." IEEE Transactions on Neural Networks and Learning Systems 35.4 (2022): 5355-5366.
[53] Hatamizadeh, Ali, et al. "Unetr: Transformers for 3d medical image segmentation." Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2022.
[54] Landman, Bennett, et al. "Multi-atlas labeling beyond the cranial vault." URL: https://www. synapse. Org (2015).
[55] Wu, Yixuan, et al. "D-former: A u-shaped dilated transformer for 3d medical image segmentation." Neural Computing and Applications 35.2 (2023): 1931-1944.
[56] Qiao, JunBo, et al. "MBUTransNet: multi-branch U-shaped network fusion transformer architecture for medical image segmentation." International Journal of Computer Assisted Radiology and Surgery 18.10 (2023): 1895-1902.
[57] Zhu, Hongyu & Xie, Chao & Fei, Yeqi & Tao, Huanjie. (2021). Attention Mechanisms in CNN-Based Single Image Super-Resolution: A Brief Review and a New Perspective. Electronics. 10. 1187. 10.3390/electronics10101187.
[58] Kumar, Neeraj, et al. "A multi-organ nucleus segmentation challenge." IEEE transactions on medical imaging 39.5 (2019): 1380-1391.
[59] Chen, Jieneng, et al. "Transunet: Transformers make strong encoders for medical image segmentation." arXiv preprint arXiv:2102.04306 (2021).
[60] Zhao, Jiamei, Dikang Wu, and Zhifang Wang. "TU-Net: U-shaped Structure Based on Transformers for Medical Image Segmentation." International Conference of Pioneering Computer Scientists, Engineers and Educators. Singapore: Springer Nature Singapore, 2022.
[61] Zhang, Jianming & Xing, Zi & Wu, Mingshuang & Gui, Yan & Zheng, Bin. (2024). Enhancing low-light images via skip cross-attention fusion and multi-scale lightweight transformer. Journal of Real-Time Image Processing. 21. 10.1007/s11554-024-01424-w.
[62] Wong, Ken CL, Hongzhi Wang, and Tanveer Syeda-Mahmood. "HartleyMHA: Self-attention in Frequency Domain for Resolution-Robust and Parameter-Efficient 3D Image Segmentation." International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer Nature Switzerland, 2023.
[63] Mehran, Meer & Pittie, Tanu & Chakraborty, Souvik & Krishnan, N M Anoop. (2022). Learning the stress-strain fields in digital composites using Fourier neural operator. iScience. 25. 105452. 10.1016/j.isci.2022.105452.
[64] Rodriguez, G. "Hartley transform: basic theory and applications in oceanographic time series analysis." WIT Transactions on Ecology and the Environment 58 (2002).
[65] Chen, Cheng, et al. "Robust multimodal brain tumor segmentation via feature disentanglement and gated fusion." Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22. Springer International Publishing, 2019.
[66] Xu, Jing. "MS-Twins: Multi-Scale Deep Self-Attention Networks for Medical Image Segmentation." arXiv preprint arXiv:2312.07128 (2023).
[67] Ouyang, Tianxiang, et al. "Rethinking U-net from an attention perspective with transformers for osteosarcoma MRI image segmentation." Computational Intelligence and Neuroscience 2022 (2022).

[68] Wang, Jinhong, et al. "Large window-based mamba unet for medical image segmentation: Beyond convolution and self-attention." arXiv preprint arXiv:2403.07332 (2024).
[69] Lin, Xian, et al. "ConvFormer: Plug-and-Play CNN-Style Transformers for Improving Medical Image Segmentation." International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer Nature Switzerland, 2023.
[70] Zhang, Z., B. Sun, and W. Zhang. "Pyramid Medical Transformer for Medical Image Segmentation. arXiv 2021." arXiv preprint arXiv:2104.14702.
[71] Oreiller, Valentin, et al. "Head and neck tumor segmentation in PET/CT: the HECKTOR challenge." Medical image analysis 77 (2022): 102336.
[72] Valanarasu, Jeya Maria Jose, et al. "Medical transformer: Gated axial-attention for medical image segmentation." Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24. Springer International Publishing, 2021.
[73] Cheng, Hao & Lian, Dongze & Deng, Bowen & Gao, Shenghua & Tan, Tao & Geng, Yanlin. (2019). Local to Global Learning: Gradually Adding Classes for Training Deep Neural Networks. 4743-4751. 10.1109/CVPR.2019.00488.
[74] Li, Shaohua, et al. "Medical image segmentation using squeeze-and-expansion transformers." arXiv preprint arXiv:2105.09511 (2021).
[75] Gonçalves, Tiago & Rio-Torto, Isabel & Luís, Teixeira & Cardoso, Jaime. (2022). A Survey on Attention Mechanisms for Medical Applications: are we Moving Toward Better Algorithms? IEEE Access. PP. 1-1. 10.1109/ACCESS.2022.3206449.
[76] Fang, Huihui, et al. "REFUGE2 challenge: A treasure trove for multi-dimension analysis and evaluation in glaucoma screening." arXiv preprint arXiv:2202.08994 (2022).
[77] Li, Ke & Wang, Di & Wang, Xu & Liu, Gang & Wu, Zili & Wang, Quan. (2023). Mixing Self-Attention and Convolution: A Unified Framework for Multi-source Remote Sensing Data Classification. IEEE Transactions on Geoscience and Remote Sensing. PP. 1-1. 10.1109/TGRS.2023.3310521.
[78] Rahman, Md Motiur, et al. "MIST: Medical Image Segmentation Transformer with Convolutional Attention Mixing (CAM) Decoder." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2024.
[79] Chen, Jia, et al. "DSNet: A dynamic squeeze network for real-time weld seam image segmentation." Engineering Applications of Artificial Intelligence 133 (2024): 108278.
[80] Khan, Tariq M., Syed S. Naqvi, and Erik Meijering. "ESDMR-Net: A lightweight network with expand-squeeze and dual multiscale residual connections for medical image segmentation." Engineering Applications of Artificial Intelligence 133 (2024): 107995.
[81] Li, Weisheng & Peng, Xiuxiu & Fu, Jun & Wang, Guofen & Huang, Yuping & Chao, Feifei. (2021). A multiscale double-branch residual attention network for anatomical–functional medical image fusion. Computers in Biology and Medicine. 141. 105005. 10.1016/j.compbiomed.2021.105005.
[82] Luo, Haozhe, Yu Changdong, and Raghavendra Selvan. "Hybrid ladder transformers with efficient parallel-cross attention for medical image segmentation." International conference on medical imaging with deep learning. PMLR, 2022.
[83] Wang, Xiyu & Guo, Pengxin & Zhang, Yu. (2023). Unsupervised Domain Adaptation via Bidirectional Cross-Attention Transformer. 10.1007/978-3-031-43424-2_19.
[84] He, Along, et al. "H2Former: An efficient hierarchical hybrid transformer for medical image segmentation." IEEE Transactions on Medical Imaging (2023).
[85] Jha, Debesh & Smedsrud, Pia & Riegler, Michael & Halvorsen, Pål & de Lange, Thomas & Johansen, Dag & Dagenborg, Håvard. (2019). Kvasir-SEG: A Segmented Polyp Dataset. 10.1007/978-3-030-37734-2_37.

[86] Zhang, Yiheng, et al. "TranSegNet: hybrid CNN-vision transformers encoder for retina segmentation of optical coherence tomography." Life 13.4 (2023): 976.
[87] Li, Zihan, et al. "Tfcns: A cnn-transformer hybrid network for medical image segmentation." International Conference on Artificial Neural Networks. Cham: Springer Nature Switzerland, 2022.
[88] Chen, Renqi, et al. "SSHNN: Semi-Supervised Hybrid NAS Network for Echocardiographic Image Segmentation." ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2024.
[89] Leclerc, Sarah, et al. "Deep learning for segmentation using an open large-scale dataset in 2D echocardiography." IEEE transactions on medical imaging 38.9 (2019): 2198-2210.