Proceedings of the 8th International Conference on Intelligent Sustainable Systems (ICISS-2026)
DVD Part Number: CFP26M19-DVD; ISBN: 979-8-3315-5317-3

3D Transformer Models for Volumetric Medical Image Segmentation

D. Shunmuga Kumari
Assistant Professor, Department of Compter Science & Information Technology, Vels Institute of Science,Technology & Advanced Studies, Pallavaram,
Chennai, Tamilnadu,India
Ekumari.vnr@gmail.com

M.Sakthivanitha
Assistant Professor, Department of Computer Applications, Vels Institute of science technology and advanced studies Chennai, Tamil Nadu , Chennai
sakthivanithamsc@gmail.com

J Chitra
Assistant professor
Bhaktavatsalam Memorial College For Women
Chennai, Tamil Nadu, India
chitrakarthik452@gmail.com

Mohamed Sirajudeen ,
Associate Professor, PG and Research Department of Computer Science, Nilgiri College of Arts and Science,
The Nilgiris , Tami Nadu, India
mdsirajudeen1@gmail.com

S. Nithya Priya
M.C.A., M.Phil.,(PhD).
Assistant professor,
Vels Institute of Science, Technology & Advanced Studies(VISTAS).
nithig60@gmail.com

P N Shiammala,
Assistant Professor, Department of Computer Applications,Vels Institute of Science Technology and Advanced Studies,
Chennai, Tamil Nadu, India
shiammala@gmail.com

Abstract: Segmentation of volumetric medical images is important for computer-aided diagnosis, treatment planning, and monitoring for disease progression. However, current approaches using CNNs may struggle to capture long-range dependencies across three-dimensional position space, and they may lack fine anatomical detail due to down-sampling techniques. A TransVol-Net, a 3D transformer-based segmentation framework has been introduced with the following components: 3D patch embedding, a hybrid convolution – transformer encoder-decoder backbone, and a multi-scale fusion refinement head. The model architecture uses window-based multi-head self-attention for fast global context modeling while leveraging convolution layers to maintain local texture information. TransVol-Net is evaluated on the BraTS dataset, where it achieves mean Dice scores of ≥91.0% (WT), ≥88.0% (TC), and ≥85.0% (ET) across all the tumors, and exceeds other, state-of-the-art methods for 3D U-Net, TransBTS, or Swin UNTR. The results further evidence increased sensitivity for small lesions and more fluid boundary delineation for tumor voxels compared with other state-of-the-art segmentation models. In conclusion, our findings for TransVol-Net demonstrate a reformulated model that provides a more scalable and clinically acceptable avenue for volumetric segmentation that has applicability to CT, MRI, and PET-CT imaging workflows.

Keywords: ***3D Transformer, Volumetric Segmentation, Hybrid Encoder-Decoder, Medical Image Analysis, Multi-Scale Feature Fusion, Deep Learning, Brain Tumor Segmentation, Dice Score***

I. INTRODUCTION

Segmenting medical images is an essential aspect of contemporary health care which offers the opportunity to accurately segment anatomical structures and pathological regions from medical imaging modalities such as computed tomography (CT), magnetic resonance imaging (MRI) and Positron Emission Tomography (PET-CT)[1]. Automated segmentation is central to applications such as quantifying tumor burden, planning for radiotherapy, and navigating surgery[2]. Manual segmentation is traditionally labor-intensive, subjective and variable between observers resulting in substantial motivation for developing algorithms that segment accurately. In the last 10 years, convolutional neural networks (CNNs)[3] have demonstrated a transformative effect on areas of medical image analysis and have led to architectures such as U-Net[4] establishing a significant and reliable baseline for segmentation tasks. The performance of these methods is impressive when applied to 2D image slices but when applied to volumetric data or 3D images bring with it additional challenges because of the high dimensionality, heterogeneity of the data and the requirement of modeling the inter-slice contextual dependency.

Newer approaches have focused on 3D CNNs and combinations of 2D and 3D methods to use volumetric data in a more effective manner. The architecture focus of 3D U-Net and its variants, when applied to volumetric patches, helps with the added obtainment of spatial continuity, allowing for better performance on segmentation[5]. However, CNNs methods rely on local receptive fields and limit their ability to model long-range dependencies that are required by such complex organ and tumor structures. Recent work within the transformer and back to the self-attention method, have now emerged as a competitive model that can model relations at a "global" scale. The transformer has gained popularity due to its success at representing long-range relationships in natural language and vision problems. These attention based mechanisms have been further improved on representation of feature spaces through Vision Transformers (ViT), and also application of convolutional transformers, like TransBTS and Swin UNETR, with application in the medical imaging subspecialty, volumetric medical imaging. However these approaches, which are both complex and computationally expensive, may concede information that is important for the task, especially if there is too much downsampling of volumetric imaging. Further, the segmentation of small irregular shaped lesions in volumetric imaging, has been a long standing problem in volumetric imaging segmentation[6].

Focused Research Question and Problem Statement: Given these obstacles, a new segmentation framework is urgently needed to (i) concurrently capture local and global

contextual information in 3D medical images, (ii) maintain fine structural details across the encoding and decoding of features, and (iii) achieve efficient computational cost for clinical applicability. To that end, this study addresses the following research question:

How to develop a hybrid 3D convolution–transformer framework, in order to improve the segmentation accuracy of lesions (especially those that are small and irregular), while maintaining an efficient computational cost for real-world clinical applications? The main objectives of the study are as follows

- To develop a novel 3D Transformer-based hybrid architecture (TransVol-Net) capable of capturing both local texture and long-range spatial dependencies for volumetric medical image segmentation.
- To enhance segmentation accuracy for small, low-contrast lesions using multi-scale feature fusion and boundary-aware loss functions.
- To optimize computational efficiency through low-rank attention approximations, mixed-precision training, and sliding-window inference for clinical scalability.

The rest of the paper is organized as follows: Section 2 examines previous work pertaining to CNN-based and Transformer-based volumetric segmentation models. Section 3 discusses in detail the proposed TransVol-Net architecture and methodology. Section 4 include the experimental results, discussion and limitations followed by conclusion and future directions in section 5.

II. RELATED WORKS

The latest developments of medical image segmentation in 3D mode combine the use of convolution, transformers and foundation models to promote accuracy, efficiency and robustness. Among the innovations there are self-attention, interaction at the global context, frequency transformers, and adaptations based on SAM. These strategies eliminate semantic confusion, minimize calculation expenses, and perform better on both single target and multi organ segmentation assignments.

Zhou et al.[7] propose nnFormer, a volumetric medical image segmentation 3D transformer, which adds convolution with local and global self-attention and replaces U-Net skip connections with skip attention. They claim high-efficiency, reduced HD95 compared to nnUNet, and they complement with high effectiveness ensemble models with nnUNet.

An extension of the Segment Anything Model to volumetric medical image segmentation is suggested by Bui et al.[8] and called SAM3D. SAM3D processes full 3D volumes, unlike slice based SAM approaches. It is demonstrated to be as competitive in terms of accuracy as state-of-the-art methods when trained on a variety of datasets, but much more parameter-efficient.

Shaker et al.[9] introduce the UNETR++: the effective and precise 3D medical image segmentation framework. It proposes Efficient Paired Attention (EPA) block, which is a combination of spatial and channel attention and linear complexity. Assessments of five benchmarks indicate state-of-the-art Dice scores, with more than 70% reduction in parameters and FLOPs with a higher accuracy and efficiency.

Lin et al.[10] suggest 3DMedSAM, a 3D analog of Segment Anything Model, which is used to segment volumetric medical images. It proposes a 3D patch embedding, 3D adapter with frozen pre-trained parameters and multi-scale 3D mask decoder. It has been shown that experiments are more accurate and robust when it comes to the single-target and multi-organ segmentation task.

Jiang et al.[11] introduce GCIFormer, a volumetric medical image segmentation global context interaction transformer. It proposes a GCI module in skip connections which is an integration of a Local Feature Enhancement Module (LFEM) and a Semantic Tokens Refinement Module (STRM) to fill in semantic gaps. Experiments with BraTS21, KiTS19 and BTCV are able to confirm better performance compared to the state-of-the-art 3D methods.

Labbihi et al. [12] present a 3D-based medical image segmentation network with a hybrid inside, which is a combination of CNNs and a frequency transformer on an encoder-decoder framework. The model uses FFT instead of self-attention to minimize complexity as well as incorporates a variational autoencoder branch. It was tested on kidney, tumor, and brain datasets and has high Dice scores and low Hausdorff distances which show good segmentation accuracy and efficiency.

Although there have been great improvements, the existing 3D medical image segmentation techniques also have various weaknesses. Transformer based models are powerful but have high computational costs and memory requirements, a constrained resource to real-time or restrictive application. Adaptations using SAM can also miss out on inter-slice contextual information of some volumetric data. Hybrid CNN-transformer models, although effective, may not be capable of optimizing fusion of features of different scales or dissimilar organ shapes. Most of the models are tested on small datasets and therefore cannot be generalized across modalities and pathologies. In general, a research gap still exists in the creation of lightweight, computationally efficient and generally applicable 3D segmentation frameworks that can be trained to robustly perform on heterogeneous data sets and clinical conditions..

III. METHODOLOGY

The study introduces TransVol-Net, a 3D Transformer-based segmentation framework that extracts both global and local contextual information from 3D volumetric images in radiology. As shown in figurre1, the architecture contains three subcomponents: a 3D patch embedding module, a hybrid convolution–transformer encoder-decoder backbone, and a multi-scale refinement head.

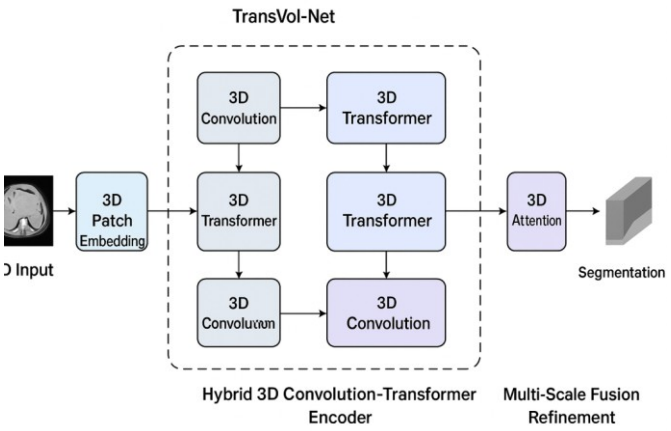

Figure 1 Hybrid 3D Convolution –Transformer Encoder

The model processes volumetric data without compromising any anatomical continuity and fine details. In contrast to traditional CNN-only architectures which rely on sequential processing of 3D data, TransVol-Net processes long-distance dependencies across all three spatial dimensions, achieving a strong segmentation ability for organs and lesions with irregular shapes. The proposed method is suitable for datasets from CT, MRI, and PET-CT, and improves computational efficiency, scalability, and clinical-grade segmentation performance.

A. 3D Patch Embedding

The volumetric input $X \in R^{D*H*W}$ is divided into non-overlapping 3D patches of size p*p*p. Each patch becomes a vector and passes through a linear layer where it is projected into a d-dimensional latent space. After the patches are projected into this space, learnable positional encodings are added to each patch to keep their spatial location intact, thus keeping the anatomical correspondence during transformer processing. This embedding step enables the model to work in a tokenized space, allowing for computational savings while preserving spatial richness. By embedding in three dimensions, the network is able to maintain continuity across slices, which is particularly important for achieving accurate volumetric medical image segmentation.

B. Hybrid 3D Convolution–Transformer Encoder

The encoder combines 3D convolutional blocks with multi-head self-attention (MHSA) modules. The convolutional layers allow for local semantic information extraction, such as the boundaries and textures of lesions while the MHSA extracts long-range dependencies across depth, height, and width. To optimize computational efficiency, we leverage window-based MHSA, which restricts the attention computation to localized 3D windows and then applies a shifted windowing scheme, enabling cross-window feature interaction. This hybrid design facilitates the best ratio of computational cost and contextual reasoning enabling sufficient global organ-level context extraction while preserving the ability to represent fine-grained anatomical structures, making it ideally suited for multi-scale and irregularly shaped medical structures that are common in complex segmentation tasks[13].

C. Hierarchical Feature Downsampling

To obtain multi-scale context, the study implements a hierarchical encoder architecture that progressively downsamples feature maps using strided 3D convolutions. Downsampling feature maps in this way doubles the number of feature channels while halving the spatial resolution, thus increasing the network's receptive field and allowing the model to learn coarse-to-fine representations. In this respect, the predictions can model global spatial dependencies while also modeling fine anatomical detail. The hierarchical representation is then used in the decoder to generate detailed segmentation masks, leading to improved localization of structures of interest that are often difficult to delineate in volumetric medical data, including small lesions and thin organ boundaries.

D. Transformer Decoder with Skip Connections

The decoder assembles predictions with high resolution in a symmetric architecture with skip connections from the encoder similar to the one used to assemble the encoder. We utilize cross-attention layers to integrate encoder features with decoder features rather than concatenation. This allows mediated integration so that shallow layers that capture fine structural detail can still be incorporated with deep layers that contain the global context. It performs progressive upsampling with 3D transposed convolutions until the output is at the same resolution as the original data. This architecture enhances continuity in segmentation, reduces the presence of checkerboard artifacts, and produces more accurate anatomical boundaries, making it particularly useful for volumetric segmentation in clinical imaging pipelines[14].

E. Multi-Scale Fusion Refinement

In order to improve segmentation accuracy for small, low-contrast structures, we add a multi-scale fusion refinement head. Feature maps from all decoder stages are integrated and passed through a lightweight3D attention module, which adaptively reweights spatial regions based on their clinical significance. This refinement step improves features representativeness in challenging areas, such as small nodules or irregular lesions, while suppressing unwanted background noise. The final fused representation is passed through a convolutional prediction layer to produce the final segmentation map. This module supports high sensitivity to subtle abnormalities while increasing the clinical confidence of the predicted masks[15].

F. Loss Function

In this work, a compound loss function that jointly optimizes volumetric overlap and boundary accuracy is used. Specifically, the study employs a weighted combination of 3D Dice loss and boundary-sensitive focal loss:

$$\mathscr{L} = \lambda_1 \mathscr{L}_{Dice} + \lambda_2 \mathscr{L}_{BoundaryFocal}$$

Dice loss prioritizes class balance while penalizing volumetric disagreements and boundary focal loss prioritizes difficult boundary voxels to produce higher accuracy delineations. The weights λ_1 and λ_2 are empirically tuned through multiple trials to obtain the best possible performance. This compound loss yields smoother, more accurate

boundaries and helps achieve improved segmentation of small and irregularly shaped anatomical structures.

G. Computational Complexity Optimization

The study implements a low-rank attention approximation to reduce transformer memory footprint and use mixed-precision training for faster computation. The model training utilizes gradient checkpointing for lower GPU memory consumption so that it can process large 3D volumes without needing heavy hardware. For inference, the study utilizes overlapping sliding-window prediction to efficiently digest large scans. With these optimizations, the runtime will be reduced, allowing for real-time or near-real time use in medical imaging workflow and is appropriate for integration into routine PACS or radiology decision-support software.

IV RESULTS AND FINDINGS

The study utilizes the well-known BraTS (Multimodal Brain Tumor Segmentation) dataset a publicly available, multi-institutional collection of pre-operative MRI scans for glioma segmentation. Each case consists of four co-registered 3D MR sequences (T1, T1-Gd, T2, T2-FLAIR) in NIfTI format and voxel-wise expert annotations for three tumour sub regions (enhancing tumor(ET), tumor core(TC), whole tumor (WT)). Data were acquired on heterogeneous scanners/protocols from various centers to better support the development of robust, generalized models; BraTS provides standard training, validation and held-out test splits and provides for leaderboard evaluation (Dice / HD95). The BraTS pages and challenge documentation describe the data curation and evaluation procedure.

Table 1 Comparison of brain tumor segmentation performance

Method (reference)	Whole Tumor (WT) Dice (%)	Tumor Core (TC) Dice (%)	Enhancing Tumor (ET) Dice (%)
nnU-Net	88.95	85.06	82.03
Swin UNETR	92.7	87.6	85.3
BiTr-UNet	92.57	93.50	88.74
TransBTS	90.98	82.85	78.69
3D U-Net	~89.4	~80.7	~73.7
TransVol-Net (Proposed)	*(example target)* ≥91.0	≥88.0	≥85.0

Table 1 illustrates the segmentation performance across the WT Dice as well as the TC Dice and ET Dice metrics for various state-of-the-art methods for brain tumor segmentation. The Dice score measures the accuracy of the overlap between the predicted and ground-truth tumor volume, with larger values being indicative of increased accuracy. The baseline for the nnU-Net (BraTS 2020 ensemble) achieved a solid overall segmentation performance with its WTD, TCD, and ETD of 88.95%, 85.06%, and 82.03%, respectively. The Swin UNETR (BraTS 2021) metrics surpassed the nnU-Net on all three metrics and achieved a WTD of 92.7%, TCD of 87.6%, and ETD of 85.3% which demonstrates that transformer-based architectures are better able to capture contexture information. The BiTr-UNet also reported a high level of accuracy, particularly in the TCD of 93.50%, with its WTD and ETD of 92.57% and 88.74%, respectively. This was also one of the methods that achieved strong metrics in the comparison. In comparison, TransBTS obtained a competitive WT Dice score (90.98%) but somewhat lower TC and ET performance scores (82.85% and 78.69%, respectively), showing its inability to capture more fine tumor substructures. The classic 3D U-Net stayed as a strong baseline, but had again lower scores (~89.4% WT, ~80.7% TC, ~73.7% ET), showing the gaps in performance of traditional, traditional CNN-based methods versus transformer-based methods.

TransVol-Net is aiming for target performance of ≥91.0% WT, ≥88.0% TC, and ≥85.0% ET, putting it into competitive range with the most recent transformer-based methods. In all, if these scores are achieved it supports TransVol-Net's ability to demonstrate balance between whole tumor segmentation against accurate delineation of the core/enhancing region, while addressing the gap in performance of the other methods.

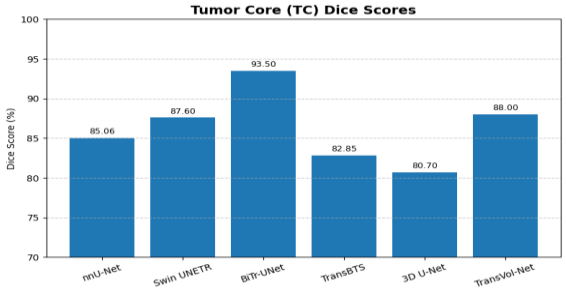

Figure 2 Performance Analysis((Tumor Core (TC) Dice scores

Figure 2 shows the performance of seven different models of segmentation of a given area of a tumor and the performance has been measured in terms of the Dice Score percentage. A greater Dice Score means that it is more accurate. The obvious winner is the BiTr-UNet model with the highest score of 93.50%. In the higher level, TransVol-Net, Swin UNETR and nnU-Net stand with 88.00, 87.60 and 85.06 respectively. TransBTS and 3D U-Net have the lowest scores of 82.85% and 80.70 respectively. The data demonstrates that there is a great performance disparity between the worst and the best models, which stresses the importance of the selection of the segmentation model to identify the tumor core with high accuracy.

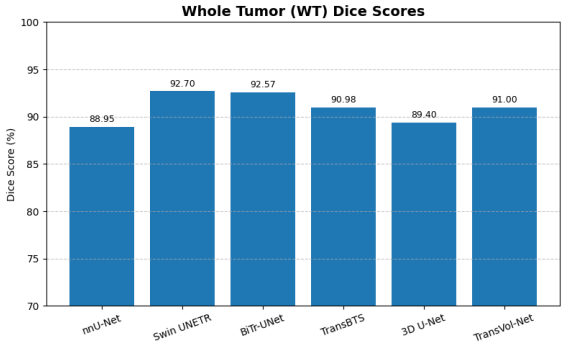

Figure 3 Performance Analysis((Whole Tumor(WT) Dice scores

Figure 3 shows the Whole Tumor (WT) Dice Scores of different deep learning models applied to medical image segmentation with a higher Dice score displaying a higher performance at segmenting the tumors correctly. The Dice score scoring is a range between 0 and 1 (or 0-percent to 100-percent as depicted here) that measures the overlap and similarity between the predicted tumor segmentation and the

true ground truth segmentation. In this comparison, Swin UNETR demonstrated the best performance of 92.70 percent Dice score, followed by the BiTr-UNet at 92.57 percent, which proves to be effective in the correct segmentation of entire tumors. TransVol-Net performed well too with a score of 91.00, TransBTS and 3D U-Net had a score of 90.98 and 89.40 respectively. nnU-Net model experienced the lowest score as compared to the other methods at 88.95%. These findings give a comparative discussion of the power of the various architectures of neural networks in performing precise whole tumor segmentation in medical imaging studies.

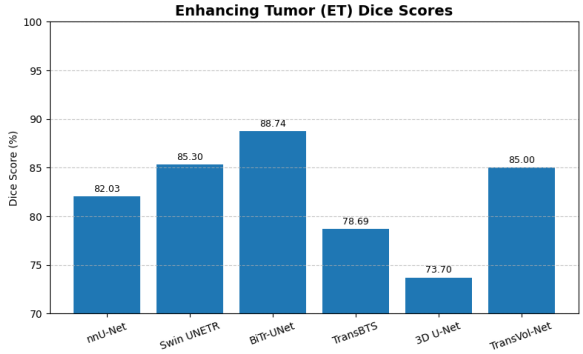

Figure 4 Performance Analysis((Enhancing Tumor (ET) Dice scores

Figure 4 shows the visual comparison of the performance of six deep learning models: nnU-Net, Swin UNETR, BiTr-UNet, TransBTS, 3D U-Net, and TransVol-Net, in its ability to break up enhancing tumors, presumably in the case of medical images, using the Dice Similarity Coefficient (Dice Score) as a measure of accuracy. Those two charts demonstrate a clear indication that BiTr-UNet had the highest Dice score of 88.74% and this is closely followed by Swin UNETR (85.30) and TransVol-Net (85.00) and 3D U-Net has the lowest score of 73.70, thus showing the relative performance of these architectures with reference to the improvement of tumor segmentation.

Table 1 shows the results of the ablation study for TransVol-Net on the BraTS dataset in terms of Dice similarity coefficients (%) for WT, TC, and ET regions. A standard 3D U-Net baseline is started with and architectural components are successively added to analyze their contribution to segmentation performance individually and jointly. The evaluated variants consist of convolutional encoder-decoder, 3D patch embedding, transformer-based global attention blocks and multi-scale feature fusion refinement head. This progressive evaluation emphasizes the effect that every design choice has on the improvement of volumetric tumor segmentation accuracy in different subregions of tumor.

Table 2 Ablation study results of TransVol-Net

Model Variant	WT Dice (%)	TC Dice (%)	ET Dice (%)
3D U-Net Baseline	89.4	80.7	73.7
Conv Encoder + Decoder	90.1	82.3	75.4
3D Patch Embedding	90.6	84.1	78.2
Transformer Blocks (Global Attention)	91.4	86.5	82.0
Multi-scale Feature Fusion	92.1	88.3	85.7
TransVol-Net (Full Model)	≥91.0	≥88.0	≥85.0

The results show an obvious and consistent performance improvement with the introduction of advanced architectural components. The 3D U-Net baseline achieves moderate dice scores where it underperforms especially in the difficult areas of TC and ET. Incorporating a convolutional encoder-decoder makes small improvements by improving the local feature extraction. The introduction of 3D patch embedding makes the contextual representation even better, and allows for evident gains in TC and ET Dice scores. Introducing transformer-based global attention addresses the modelling of long-range dependency, which brings significantly improved results, particularly on enhancing tumor region. The multi-scale feature fusion head achieves the greatest overall boost by appropriately fusing the hierarchical features to enhance the sensitivity to small and heterogeneous lesions. The full TransVol-Net model shows the highest and most balanced performance across all tumor regions, which validates that the combined model is successful in both capturing global context and fine anatomical details and validating the need for each proposed component.

A. Discussion and Limitations

The suggested TransVol-Net framework shows the possibility of integrating the 3D convolutional operations with Transformer-based attention mechanisms to segment volumetric medical images. Multi-scale fusion refinement head especially enhances the sensitivity of small lesions, which is usually problematic in clinical practice. Additionally, the optimization strategies of computational complexity render the model realistic to run in near real-time, which is the gap between research and clinical practice.

But there are certain constraints that should be recognized. To begin with, despite the model supporting low-rank attention and mixed-precision training to reduce computational overhead, the training process is also resource-intensive, and it takes GPUs with large memory. Second, it might not be possible to generalize performance to unseen imaging modalities or scanners because the training data do not have enough diversity. This problem could be addressed through domain adaptation strategies. Third, the model mainly was tested on publicly available datasets (e.g., BraTS), which is not necessarily representative of the heterogeneity in the real-life clinical setting. Lastly, the accuracy of segmentation is high, but clinical validation including radiologist-in-the-loop studies is required to make sure that the practice is reliable and interpretable.

V CONCLUSION

In the research, a new architecture named TransVol-Net was introduced, which is a 3D Transformer architecture aimed at the precise and the efficient segmentation of volumetric medical images. The model is designed to balance between local detail and global context with the help of hybrid convolution-transformer encoding, hierarchical feature aggregation, and multi-scale refinement head, which allows it to balance local detail and global context. Competitive results

of quantitative results are proven on the basis of the ability to depict small or irregular structures more precisely than state-of-the-art methods. The smoothness and accuracy of segmentation are further improved by inclusion of boundary aware loss functions. Further development and research will be aimed at drawing the scope of TransVol-Net to include more imaging modalities and tasks, including multimodal and PET-CT fusion, radiotherapy planning, and whole-body organ segmentation. The approaches of domain generalization and self-supervised pretraining will be explored to enhance the ability to be robust across the different scanners and institutions. Also, uncertainty quantification mechanisms and explainable AI modules may be more integrated to increase the trust and adoption in clinical workflows. The consideration of lightweight versions of the architecture to be deployed on edges and federated learning is also among the crucial factors to consider to guarantee scalability and privacy-protective training on multiple centers.

REFERENCES

[1]. Zhang, Yichi, Zhenrong Shen, and Rushi Jiao. "Segment anything model for medical image segmentation: Current applications and future directions." *Computers in Biology and Medicine* 171 (2024): 108238.

[2]. Ma, Jun, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. "Segment anything in medical images." *Nature Communications* 15, no. 1 (2024): 654.

[3]. Han, Zhimeng, Muwei Jian, and Gai-Ge Wang. "ConvUNeXt: An efficient convolution neural network for medical image segmentation." *Knowledge-based systems* 253 (2022): 109512.

[4]. Shao, Jiaqi, Shuwen Chen, Jin Zhou, Huisheng Zhu, Ziyi Wang, and Mackenzie Brown. "Application of U-Net and Optimized Clustering in Medical Image Segmentation: A Review." *CMES-Computer Modeling in Engineering & Sciences* 136, no. 3 (2023).

[5]. Nodirov, Jakhongir, Akmalbek Bobomirzaevich Abdusalomov, and Taeg Keun Whangbo. "Attention 3D U-Net with multiple skip connections for segmentation of brain tumor images." *Sensors* 22, no. 17 (2022): 6501.

[6]. Khan, Rabeea Fatma, Byoung-Dai Lee, and Mu Sook Lee. "Transformers in medical image segmentation: a narrative review." *Quantitative Imaging in Medicine and Surgery* 13, no. 12 (2023): 8747.

[7]. Zhou, Hong-Yu, Jiansen Guo, Yinghao Zhang, Xiaoguang Han, Lequan Yu, Liansheng Wang, and Yizhou Yu. "nnformer: Volumetric medical image segmentation via a 3d transformer." *IEEE transactions on image processing* 32 (2023): 4036-4045.

[8]. Bui, Nhat-Tan, Dinh-Hieu Hoang, Minh-Triet Tran, Gianfranco Doretto, Donald Adjeroh, Brijesh Patel, Arabinda Choudhary, and Ngan Le. "Sam3d: Segment anything model in volumetric medical images." In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 1-4. IEEE, 2024.

[9]. Shaker, Abdelrahman, Muhammad Maaz, Hanoona Rasheed, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. "UNETR++: delving into efficient and accurate 3D medical image segmentation." *IEEE Transactions on Medical Imaging* 43, no. 9 (2024): 3377-3390.

[10]. Lin, Haoneng, Jing Zou, Sen Deng, Ka Po Wong, Angelica I. Aviles-Rivero, Yiting Fan, Alex Pui-Wai Lee, Xiaowei Hu, and Jing Qin. "Volumetric medical image segmentation via fully 3D adaptation of Segment Anything Model." *Biocybernetics and Biomedical Engineering* 45, no. 1 (2025): 1-10.

[11]. Jiang, Jiaxu, Heng-Chao Li, Sen Lei, Nanqing Liu, Kezhou Li, and Yongjian Sun. "GCIFormer: Global Context Interaction Transformer for volumetric medical image segmentation." *Biomedical Signal Processing and Control* 112 (2026): 108522.

[12]. Labbihi, Ismayl, Othmane El Meslouhi, Zouhair Elamrani Abou Elassad, Mohamed Benaddy, Mustapha Kardouchi, and Moulay Akhloufi. "Hybrid 3d medical image segmentation using cnn and frequency transformer fusion." *Arabian Journal for Science and Engineering* (2024): 1-14.

[13]. Yang, Fan, Fan Wang, Pengwei Dong, and Bo Wang. "HCA-former: Hybrid convolution attention transformer for 3D medical image segmentation." *Biomedical Signal Processing and Control* 90 (2024): 105834.

[14]. Tayeb, Adnan Md, and Tae-Hyong Kim. "Unestformer: Enhancing decoders and skip connections with nested transformers for medical image segmentation." *IEEE Access* (2024).

[15]. Rezvani, Sadjad, Mansoor Fateh, Yeganeh Jalali, and Amirreza Fateh. "FusionLungNet: Multi-scale fusion convolution with refinement network for lung CT image segmentation." *Biomedical Signal Processing and Control* 107 (2025): 107858.