SegFormer3D: an Efficient Transformer for 3D Medical Image Segmentation

Shehan Perera*, Pouyan Navard*, Alper Yilmaz
Photogrammetric Computer Vision Lab, The Ohio State University
{perera.27, boreshnavard.1, yilmaz.15}@osu.edu

Abstract

The adoption of Vision Transformers (ViTs) based architectures represents a significant advancement in 3D Medical Image (MI) segmentation, surpassing traditional Convolutional Neural Network (CNN) models by enhancing global contextual understanding. While this paradigm shift has significantly enhanced 3D segmentation performance, state-of-the-art architectures require extremely large and complex architectures with large scale computing resources for training and deployment. Furthermore, in the context of limited datasets, often encountered in medical imaging, larger models can present hurdles in both model generalization and convergence. In response to these challenges and to demonstrate that lightweight models are a valuable area of research in 3D medical imaging, we present SegFormer3D, a hierarchical Transformer that calculates attention across multiscale volumetric features. Additionally, SegFormer3D avoids complex decoders and uses an all-MLP decoder to aggregate local and global attention features to produce highly accurate segmentation masks. The proposed memory efficient Transformer preserves the performance characteristics of a significantly larger model in a compact design. SegFormer3D democratizes deep learning for 3D medical image segmentation by offering a model with 33× less parameters and a 13× reduction in GFLOPS compared to the current state-of-the-art (SOTA). We benchmark SegFormer3D against the current SOTA models on three widely used datasets Synapse, BRaTs, and ACDC, achieving competitive results. Code: https://github.com/OSUPCVLab/SegFormer3D.git

1. Introduction

The emergence of deep learning in healthcare has been transformative, offering an unprecedented capacity to learn and analyze complex medical data patterns. A fundamental task in medical image analysis is 3D volumetric image segmentation that is crucial for applications such as tumor and multi-organ localization in diagnosis and treatment. The conventional approach involves employing an encoder-decoder architecture [18, 22], where the image is first transformed into a low-dimensional representation, and then the decoder maps the representation to a voxel-wise segmentation mask. However, these architectures struggle to generate accurate segmentation masks due to their limited receptive field. Recently, Transformer-based techniques have demonstrated superior segmentation performance owing to the ViT's ability to utilize attention layers for capturing global relationships [11, 31]. This stands in sheer contrast to CNNs which exhibits local inductive bias properties.

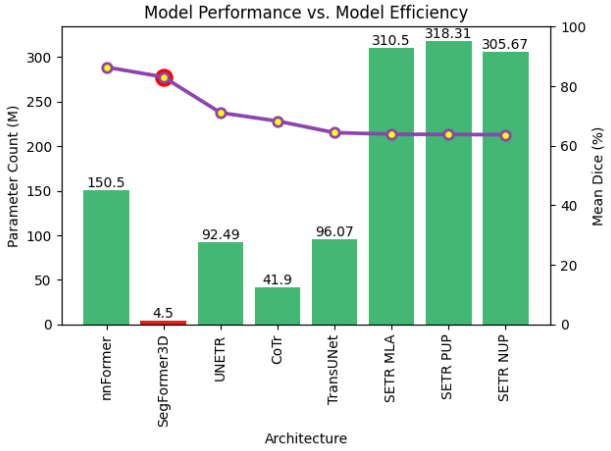

Fig. 1. **Parameter Count *vs*. Performance on BraTs** We compare Segformer3D to existing 3D volumetric image segmentation architectures evaluating model performance with respect to parameter count. The green bars represent model parameters while the purple plot shows the mean dice performance for each architecture. We demonstrate that at 4.5 million parameters Segformer3D is a highly competitive lightweight architecture for 3D medical image segmentation.

Following the seminal work of TransUnet[5] and UNETR [11], a large body of research in the medical community has been dedicated to designing Transformer based architectures that take advantage of the strong encoding capability of ViTs and the feature refinement capability of CNNs in the decoding stage. For example, [10, 11, 31] combined localized receptive field of convolutions and global attention. Despite their advantages, ViTs fail to match the generalization capabilities of CNNs when trained from scratch

*These authors contributed equally to this work

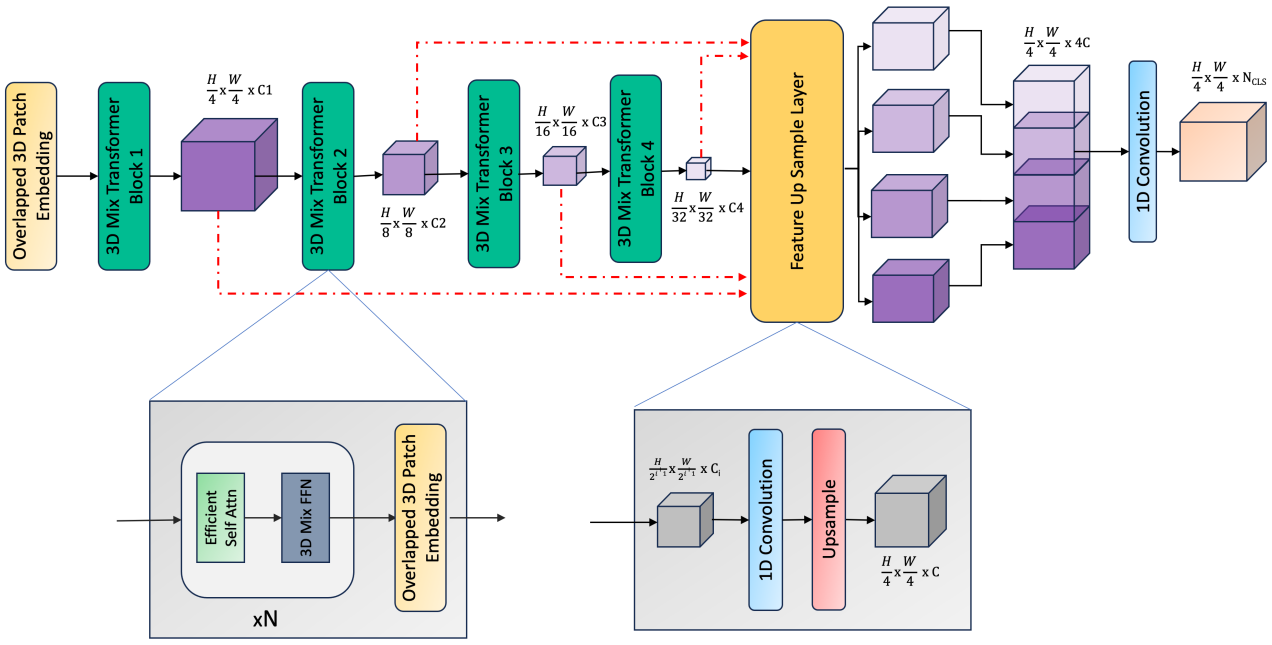

Fig. 2. Segformer3D Overview: The model input is a 3D volume $\mathbb{R}^{D \times C \times H \times W}$. We extract multiscale volumetric features using a 4 stage hierarchical Transformer. An all-MLP decoder then upsamples and aggregate local and global attention features from the encoding stage to generate the final segmentation mask.

on small-scale datasets, and because of lack of inductive bias often depends on large scale datasets for pretraining [7] which are not commonly available in medical image domain. Additionally, computational efficiency of the ViTs are bounded by number of floating point operation and element wise functions in the multi head self attention block [17]. This issue is much more prominent in 3D medical imaging tasks because of the fact that the length of the converted sequence of the 3D volumetric input is considerably long. Furthermore, medical imaging data frequently exhibits repetitive structures [6], suggesting that it can be compressed, a consideration often overlooked by the 3D SOTA ViT architectures in the medical domain.

This paper presents SegFormer3D, a volumetric hierarchical ViT, that extends [26] to 3D medical image segmentation tasks. Unlike vanilla ViT[7] which renders feature maps on a fixed scale, Segformer3D encodes feature maps at different scales of the input volume following the Pyramid Vision Transformer [25]. Our design enables the Transformer to capture a variety of coarse to fine-grained features of the input. SegFormer3D also utilizes an efficient self-attention module [25] that compresses the embedded sequence to a fixed ratio to significantly reduce model complexity without sacrificing performance Figure 1. Additionally, SegFormer3D utilizes the overlapping patch embedding module used in [26] that preserves the local continuity of the input voxels. This embedding uses a positional-free encoding [14] that prevents accuracy loss when there is a resolution mismatch during training and inference, which is common in medical image segmentation. To efficiently generate a high-quality segmentation mask, SegFormer3D uses an all-MLP decoder introduced in [26]. Comprehensive experiments on three benchmark datasets-Synapse[15], ACDC[1], and BRaTs[20]—validate the qualitative and quantitative effectiveness of SegFormer3D. Our contributions can be summarized as:

- We introduce a lightweight memory efficient segmentation model that preserves the performance characteristics of larger models for 3D medical imaging.
- With 4.5 million parameters and 17 GFLOPS, Segformer3D presents a $34\times$ and $13\times$ reduction in parameter count and model complexity vs SOTA.
- We showcase highly competitive results without pretraining, emphasizing the generalization capabilities of lightweight ViTs and that exploring architectures like Segformer3D is a valuable research area in medical imaging.

2. Related Work

Following the introduction of Unet[22], numerous approaches have been proposed for medical image analysis such as Dense-unet[2] and deep-supervised CNN [32]. Unet has also been extended to 3D medical image analysis, for instance, 3D-Unet[6], V-net[21], nn-Unet[13] and [8, 9, 23]. Researchers also designed hierarchical architectures to capture contextual information. In [21], Milletari et al. downsampled the volume to a lower resolution to preserve beneficial image features using V-net. Cicek et al.[6] replaced the 2D to 3D convolutions in 3D-unet. Isensee et al.[13] proposed the nn-Unet generalized segmentation architecture that can extract features at multiple scales. In [16], PGD-UNet uses deformable convolution to deal with irregular organ shapes and tumors for medical image segmentation.

Several recent papers have studied Transformer-convolution architectures such as TransUnet[5], Unetr[11], SwinUnetr[10], TransFuse[29], nnFormer[31]. TransUnet[5] combines Transformers and U-Net to encode image patches and decode through high resolution upsampled CNN features for localization. Hatamizadeh et al. [11] present UNETR, a 3D model merging the long-range spatial dependencies characteristic of Transformers with the inherent CNN inductive biases in a "U-shaped" encoder-decoder structure. In UNETR, Transformer-blocks encode features that capture consistent global representations and are subsequently integrated across various resolutions within a CNN-based decoder. Zhou et al.[31] present nnFormer, a method derived from the Swin-UNet[3] architecture. Wang et al.[24] proposed TransBTS which uses a regular convolutional encoder-decoder architecture and a Transformer layer as the bottleneck.

3. Method

The adoption of Transformers has greatly improved the performance of volumetric medical image segmentation. However, current high-performing architectures prioritize over-parameterization for model performance, sacrificing efficiency. To demonstrate the benefits of lightweight and efficient Transformers without compromising on performance, we introduce Segformer3D. With **4.5 million** parameters and **17 GFLOPS** we show a reduction of **34**× and **13**× in parameter count and complexity showcasing the significance of the proposed architecture in 3D medical image segmentation 1.

Encoder: Using 3D medical images within the Transformer framework results in long sequence lengths which increases the computational complexity of the model. For example, a standard 3D MRI volume with dimensions of 128^3 results in a sequence length of 32,768, whereas a typical 2D RGB image with dimensions of 256^2 yields a sequence length of 256. Our hierarchical Transformer incorporates three key elements to improve computational efficiency and reduce the total parameter count while maintaining the SOTA level performance. First, we incorporate overlapped patch merging to overcome neighborhood information loss during the voxel generation process. This technique, in contrast to the patching mechanism seen in ViT[7], allows the model to better understand the transition points between the voxels and has been shown to improve overall segmentation precision[26]. Next, to address the sequence length bottleneck without compromising performance, we integrate an efficient self-attention mechanism [25]. This approach enables the model to capture long-range dependencies more effectively, promoting improved scalability and performance. Traditional self-attention takes a sequence of vectors of shape [Batch, Sequence, Features] as input and generates 3 unique projections, Query, Key and Value vectors. Once generated, the attention scores are computed as $(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_{\text{head}}}}\right)V$. Due to the operation QK^T, the computational complexity of the original segmentation process is $\mathcal{O}(n^2)$. Although this complexity can be overlooked with 2D images, with long 3D sequences, it proves to be a challenge for efficient architecture design. Efficient attention introduced in [25, 26].

Architecture	Params	GFLOPs
nnFormer[31]	150.5	213.4
TransUnet[5]	96.07	88.91
UNETR[11]	92.49	75.76
SwinUNETR[10]	62.83	384.2
Segformer3D (ours)	**4.51**	**17.5**

Table 1. Segformer3D vs SOTA in Size (M), and complexity. Segformer3D showcases a significant reduction in parameters and computational complexity without sacrificing on performance.

$$\begin{aligned} \hat{K} &= \text{Reshape}(\frac{N}{R}, C \cdot R)(K), \\ K &= \text{Linear}(C \cdot R, C)(\hat{K}), \end{aligned}$$

significantly reduces the computational complexity generated by 3D volumetric tensors from $\mathcal{O}(n^2)$ to $\mathcal{O}(n^2/r)$. We set the reduction parameter r to $4\times, 2\times, 1\times, 1\times$ in the four stages of the encoder.

Finally, our approach addresses the challenge of resizing volumetric imaging and its relation to fixed positional encoding in ViTs by adopting the mix-ffn module [26]. This module enables automatic learning of positional cues, eliminating the need for fixed encoding, ensuring superior scalability and performance.

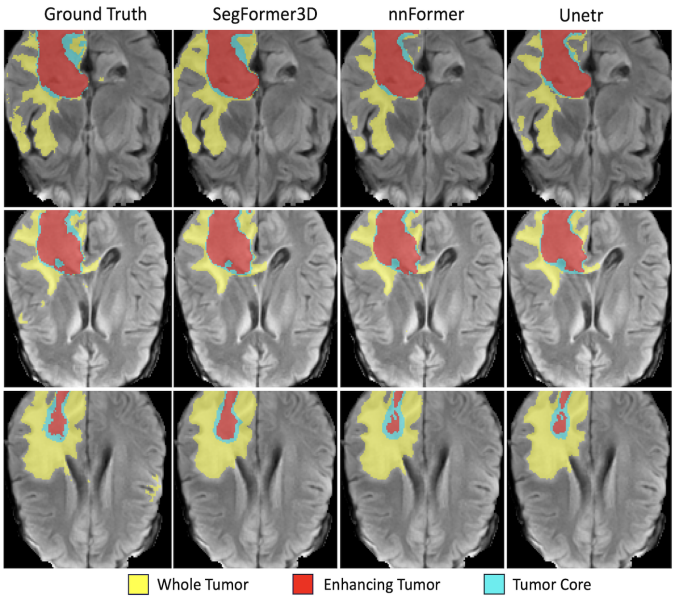

Fig. 2: Qualitative results on BRaTs. Each row is a separate frame in the MRI sequence while each column is 3D volumetric image segmentation solution. We qualitatively demonstrate highly accurate segmentation performance to SOTA methods while maintaining a lightweight and efficient architecture.

Methods	Params	Avg % ↑	Whole Tumor ↑	Enhancing Tumor ↑	Tumor Core ↑
nnFormer[31]	150.5	86.4	91.3	81.8	86.0
Ours	**4.5**	**82.1**	**89.9**	**74.2**	**82.2**
UNETR[11]	92.49	71.1	78.9	58.5	76.1
TransBTS[24]	–	69.6	77.9	57.4	73.5
CoTr[27]	41.9	68.3	74.6	55.7	74.8
CoTr w/o CNN Encoder[27]	–	64.4	71.2	52.3	69.8
TransUNet[5]	96.07	64.4	70.6	54.2	68.4
SETR MLA[30]	310.5	63.9	69.8	55.4	66.5
SETR PUP[30]	318.31	63.8	69.6	54.9	67.0
SETR NUP[30]	305.67	63.7	69.7	54.4	66.9

Table 2: BRaTs comparison table ranked based on average performance across all classes. Segformer3D is highly competitive out performing well established solutions across all categories.

Decoder: The decoding stage plays a pivotal role in medical image segmentation based on the encoder-decoder design widely adopted in the UNET based architectures[10, 11]. This framework is used in both CNN-based and Transformer-based encoders. In the context of 3D medical images, where successive 3D convolutions are often necessary for effective decoding, we instead demonstrate that the integration of linear layers is a highly effective decoding strategy for medical image segmentation. Our approach simplifies the decoding process, ensuring efficient and consistent decoding of volumetric features across diverse datasets without over-parameterization. The simple decoder process is:

$$\text{step 1:} \quad F_i = \text{Linear}(C_i, C)(F_i), \quad \forall i \tag{1}$$

$$\text{step 2:} \quad \hat{F}_i = \text{Upsample}\,(W_{4\times4})\,(\hat{F}_i), \quad \forall i \tag{2}$$

$$\text{step 3:} \quad F = \text{Linear}(4C, C)(\text{Concat}(\hat{F}_i)), \quad \forall i \tag{3}$$

$$\text{step 4:} \quad M = \text{Linear}(C, N_{\text{cls}})(F) \tag{4}$$

Similar to the skip connections introduced in UNET, features at each stage are collated, and a fixed dimensional projection is generated. Once all dimensions are standardized, we upsample each feature and perform a concatenation followed by a fusion operation. The fused features are input to a linear projection head (3D 1x1 convolutions) to generate the final segmentation masks.

4. Experimental Results

Adhering to the SOTA architecture for 3D volumetric segmentation, we utilize the same datasets and evaluation methods to ensure a fair and consistent comparison across all architectures. We train and evaluate the proposed model on three widely used datasets without the use of external data for pretraining purposes. These datasets are Brain Tumor Segmentation (BraTS) [20], Synapse Multi-Organ Segmentation (Synapse) [15], and finally the Automatic Car-

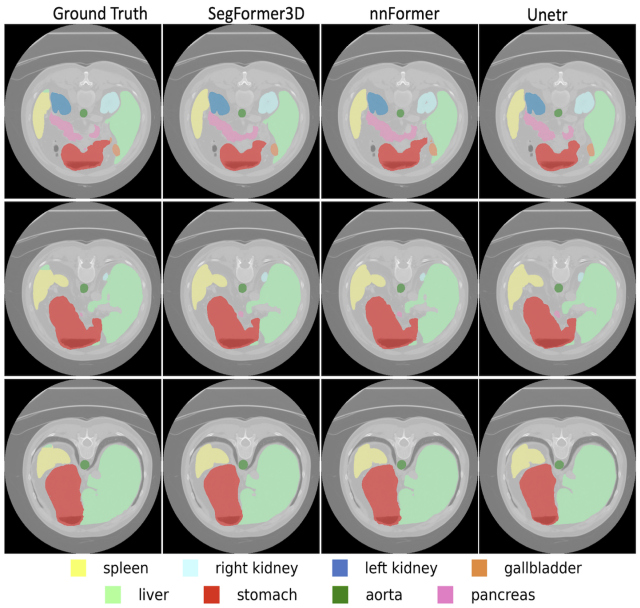

Fig. 3: Qualitative results on Synapse. Each row is a separate frame in the CT sequence while each column is different 3D volumetric image segmentation solution. Each organ mask is highlighted with a unique color code. We qualitatively demonstrate highly accurate segmentation performance compared to well estabilied SOTA methods while maintaining a lightweight design.

Methods	Params	Avg % ↑	AOR	LIV	LKID	RKID	GAL	PAN	SPL	STO
nnFormer[31]	150.5	86.57	92.04	96.84	86.57	86.25	70.17	83.35	90.51	86.83
Ours	**4.5**	**82.15**	**90.43**	**95.68**	**86.53**	**86.13**	**55.26**	**73.06**	**89.02**	**81.12**
MISSFormer[12]	–	81.96	86.99	94.41	85.21	82.00	68.65	65.67	91.92	80.81
UNETR[11]	92.49	79.56	89.99	94.46	85.66	84.80	60.56	59.25	87.81	73.99
SwinUNet[3]	–	79.13	85.47	94.29	83.28	79.61	66.53	56.58	90.66	76.60
LeVit-UNet-384[28]	52.17	78.53	87.33	93.11	84.61	80.25	62.23	59.07	88.86	72.76
TransClaw U-Net[4]	–	78.09	85.87	94.28	84.83	79.36	61.38	57.65	87.74	73.55
TransUNet[5]	96.07	77.48	87.23	94.08	81.87	77.02	63.16	55.86	85.08	75.62
R50-ViT+CUP[5]	86.00	71.29	73.73	91.51	75.80	72.20	55.13	45.99	81.99	73.95
ViT+CUP[5]	86.00	67.86	70.19	91.32	74.70	67.40	45.10	42.00	81.75	70.44

Table 3: Synapse comparisons ranked based on average performance across classes. Segformer3D is highly competitive, outperforming well-established solutions and second to only nnformer with 34x parameters.

diac Diagnosis (ACDC) [1] datasets.

All experiments, including training, real-time augmentation, and inference, were performed on a single Nvidia RTX 3090 GPU using PyTorch. Each model is trained with the same learning rate policy, which includes a learning rate warm-up stage, where we linearly increase the learning rate from $4e-6$ to $4e-4$, which is followed by a PolyLR decay strategy. The widely adopted AdamW optimizer [19] was used with a learning rate of $3e-5$. For the loss function, an equally weighted Dice-Cross Entropy Loss combination was adopted to combine the benefits of each loss function in the optimization process, improving convergence. We set the batch size to 4 and train each model for 1000 epochs, similar to the SOTA architecture. Additionally, all experiments are performed without the use of complicated pretraining efforts to showcase the proposed architectures performance on real-world medical datasets without additional data.

4.1. Results on Brain Tumor Segmentation (BraTs)

BraTs [20] is a dataset for medical image segmentation from MRI scans. The dataset contains 484 MRI images with four modalities, FLAIR, T1w, T1gd, and T2w. Data were collected from 19 institutions with ground-truth labels for three types of tumor subregions: edema (ED), enhancing tumor (ET) and nonenhancing tumor (NET). Following the same data preparation, augmentation and reporting strategies in major publications including nnFormer [31] we report our results on whole tumor (WT), enhancing tumor (ET) and tumor core (TC). In Table 3a we demonstrate that Segformer3D stands out as a strong competitive lightweight segmentation architecture against significantly larger and widely adopted CNN and Transformer architectures while maintaining 4.5 million parameters and 17.5 GFLOP computational complexity. This demonstrates the representation learning capability of the efficient self atten-

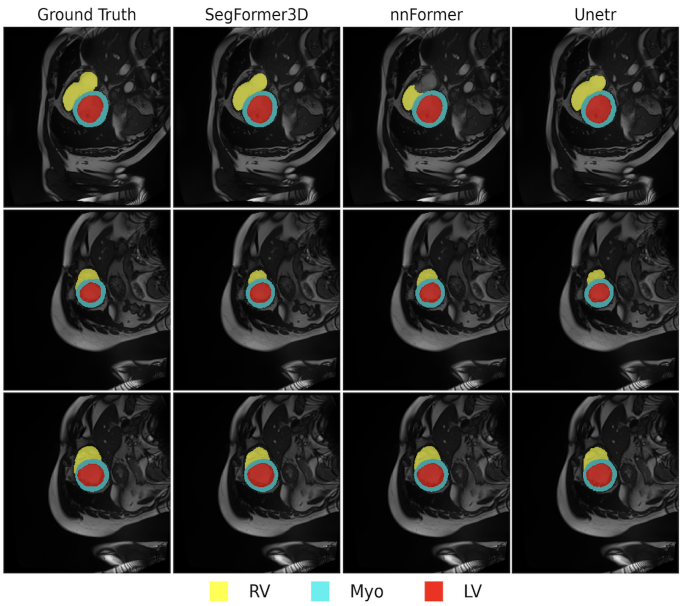

Fig. 4: Qualitative results on ACDC. Each row is a separate frame in the cine-MRI sequence while each column is different 3D volumetric image segmentation solution. We demonstrate highly accurate segmentation results to SOTA methods while maintaining a lightweight and efficient architecture.

Methods	Params	Avg % ↑	RV	Myo	LV
nnFormer [31]	150.5	92.06	90.94	89.58	95.65
Ours	**4.5**	**90.96**	**88.50**	**88.86**	**95.53**
LeViT-UNet-384 [28]	52.17	90.32	89.55	87.64	93.76
SwinUNet [3]	–	90.00	88.55	85.62	95.83
TransUNet [5]	96.07	89.71	88.86	84.54	95.73
UNETR [11]	92.49	88.61	85.29	86.52	94.02
R50-VIT-CUP [5]	86.00	87.57	86.07	81.88	94.75
VIT-CUP [5]	86.00	81.45	81.46	70.71	92.18

Table 4: ACDC comparison ranked based on average performance across classes. Segformer3D is highly competitive outperforming well established solutions and is within 1% of SOTA with 150 million parameters.

tion module over traditional ViT architectures that analyzes the whole sequence of patches without compression. Finally, we showcase highly competitive qualitative results of the proposed architecture in Figure 3a showcasing superior performance against well established architectures.

4.2. Results on Multi-Organ CT Segmentation (Synapse)

The Synapse dataset [15] provides 30 annotated CT images. We generate our results after data processing, training, and data splits defined in [31]. With a diverse set of annotations that cover multiple organs, such as spleen, pancreas, gallbladder, and others, the synapse dataset is a complex multi-class segmentation challenge. The quantitative results in Table 4a show that Segformer3D is ranked second only to the nnFormer [31] architecture with 150 million parameters. Additionally, we showcase qualitative performance results in Figure 4a where we compare highly accurate organ segmentation masks with current SOTA architectures validating the visual consistency of the proposed approach. Finally, compared to widely used architectures [10–12], Segformer3D generates competitive results with only 4.5M parameters and demonstrates that over parameterization does not lead to large performance gains, especially in data constrained situations.

4.3. Results on Automated cardiac diagnosis (ACDC)

ACDC[1] is a dataset of 100 patients used for 3D volumetric segmentation of the left (LV) and right (RV) cardiac ventricles and the myocardium (Myo)[1]. To maintain a one-to-one comparison with published research, we follow the same training and inference pipelines specified in [31] and measure segmentation accuracy using the Dice metric. Table 5a quantitatively demonstrates the proposed architecture is highly competitive against large and highly

complex solutions. The proposed model is within 1% margin of the SOTA performance with models on average 34× higher in parameter count and 13× higher in computational complexity. Comparisons of the qualitative results are visualized in Figure 5a showcasing highly competitive performance without the need for large scale pretraining on small datasets.

4.4. Conclusion

Architectures such as UNETR, TransUNet and nnFormer have revolutionized 3D volumetric medical image segmentation using the ViT framework. This paradigm shift has notably enhanced the model's contextual understanding capabilities compared to its conventional pure Convolutional Neural Network (CNN) counterparts. However, this improvement has come at the cost of a substantial increase in parameter count and model complexity, attributed to the intricate nature of the self-attention module. In addition to model size and complexities, large models prevent medical researchers with limited access to large scale compute resources from effectively training and integrating these models into their workflows. Furthermore, larger models can introduce challenges to model generalization, and convergence, especially in scenarios with limited datasets commonly seen in medical imaging. To overcome these limitations without sacrificing on performance we introduce Segformer3D a lightweight architecture that is 34× smaller in parameters and 13× less in parameters and computational complexity respectively over the state-of-the-art (SOTA) architectures. We benchmark our solution to current SOTA solutions as well as other highly cited works and we showcase that lightweight and efficient architectures can help significantly improve performance over much larger models without additional pretraining and with minimal computational resources. Finally, we assert that directing research efforts toward the development of high-performance lightweight architectures, particularly in domains with tangible real-world advantages like medical imaging, not only broadens accessibility but also promotes practical applications of such architectures in real-world scenarios.

References

[1] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging*, 37 (11), 2018. 2, 5, 6

[2] Sijing Cai, Yunxian Tian, Harvey Lui, Haishan Zeng, Yi Wu, and Guannan Chen. Dense-unet: a novel multiphoton in vivo cellular image segmentation model based on a convolutional neural network. *Quantitative imaging in medicine and surgery*, 10(6):1275, 2020. 3

[3] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*. Springer, 2022. 3, 5, 6

[4] Yao Chang, Menghan Hu, Guangtao Zhai, and Xiao-Ping Zhang. Transclaw u-net: Claw u-net with transformers for medical image segmentation. *2022 5th International Conference on Information Communication and Signal Processing (ICICSP)*, 2021. 5

[5] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv:2102.04306*, 2021. 1, 3, 4, 5, 6

[6] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, pages 424–432. Springer, 2016. 2, 3

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv:2010.11929*, 2020. 2, 3

[8] Qi Dou, Hao Chen, Yueming Jin, Lequan Yu, Jing Qin, and Pheng-Ann Heng. 3d deeply supervised network for automatic liver segmentation from ct volumes. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*. Springer, 2016. 3

[9] Eli Gibson, Francesco Giganti, Yipeng Hu, Ester Bonmati, Steve Bandula, Kurinchi Gurusamy, Brian Davidson, Stephen P Pereira, Matthew J Clarkson, and Dean C Barratt. Automatic multi-organ segmentation on abdominal ct with dense v-networks. *IEEE transactions on medical imaging*, 37(8), 2018. 3

[10] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI Brainlesion Workshop*. Springer, 2021. 1, 3, 4, 6

[11] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022. 1, 3, 4, 5, 6

[12] Xiaohong Huang, Zhifang Deng, Dandan Li, Xueguang Yuan, and Ying Fu. Missformer: An effective transformer for 2d medical image segmentation. *IEEE Transactions on Medical Imaging*, 42(5), 2023. 5, 6

[13] Fabian Isensee, Jens Petersen, Andre Klein, David Zimmerer, Paul F. Jaeger, Simon Kohl, Jakob Wasserthal, Gregor

Koehler, Tobias Norajitra, Sebastian Wirkert, and Klaus H. Maier-Hein. nnu-net: Self-adapting framework for u-net-based medical image segmentation, 2018. 3

[14] Md Amirul Islam, Sen Jia, and Neil DB Bruce. How much position information do convolutional neural networks encode? *arXiv:2001.08248*, 2020. 2

[15] Bennett Landman, Zhoubing Xu, J Igelsias, Martin Styner, T Langerak, and Arno Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, 2015. 2, 4, 6

[16] Ziqiang Li, Hong Pan, Yaping Zhu, and A Kai Qin. Pgd-unet: A position-guided deformable network for simultaneous segmentation of organs and tumors. In *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020. 3

[17] Xinyu Liu, Houwen Peng, Ningxin Zheng, Yuqing Yang, Han Hu, and Yixuan Yuan. Efficientvit: Memory efficient vision transformer with cascaded group attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14420–14430, 2023. 2

[18] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 1

[19] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017. 5

[20] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10), 2014. 2, 4, 5

[21] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016. 3

[22] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015. 1, 3

[23] Holger R Roth, Hirohisa Oda, Yuichiro Hayashi, Masahiro Oda, Natsuki Shimizu, Michitaka Fujiwara, Kazunari Misawa, and Kensaku Mori. Hierarchical 3d fully convolutional networks for multi-organ segmentation. *arXiv:1704.06382*, 2017. 3

[24] Wenxuan Wang, Chen Chen, Meng Ding, Hong Yu, Sen Zha, and Jiangyun Li. Transbts: Multimodal brain tumor segmentation using transformer. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*. Springer, 2021. 3, 4

[25] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2021. 2, 3

[26] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing*, 34, 2021. 2, 3

[27] Yutong Xie, Jianpeng Zhang, Chunhua Shen, and Yong Xia. Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. *ArXiv*, abs/2103.03024, 2021. 4

[28] Guoping Xu, Xingrong Wu, Xuan Zhang, and Xinwei He. Levit-unet: Make faster encoders with transformer for medical image segmentation. *ArXiv*, abs/2107.08623, 2021. 5, 6

[29] Yundong Zhang, Huiye Liu, and Qiang Hu. Transfuse: Fusing transformers and cnns for medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021*. Springer, 2021. 3

[30] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H. S. Torr, and Li Zhang. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 4

[31] Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnformer: Interleaved transformer for volumetric segmentation. *arXiv:2109.03201*, 2021. 1, 3, 4, 5, 6

[32] Qikui Zhu, Bo Du, Baris Turkbey, Peter L Choyke, and Pingkun Yan. Deeply-supervised cnn for prostate segmentation. In *2017 international joint conference on neural networks (IJCNN)*. IEEE, 2017. 3