Supplementary Material

Amin Ranem, Camila González, Anirban Mukhopadhyay
GRIS, Technical University of Darmstadt
Karolinenpl. 5, 64289 Darmstadt, Germany
{amin.ranem, camila.gonzalez, anirban.mukhopadhyay}@gris.informatik.tu-darmstadt.de

This supplementary material is split into two components. We first present the state-of-the-art when it comes to Transformer architectures and their application for medical image segmentation. This is used in addition to the main manuscript to better frame our contribution and to put the main manuscript in context with existing work. Secondly, we include a simple, non-CL comparison between the no new Net (nnU-Net) and our proposed architecture from the manuscript.

1. Related Work

Transformers as presented in Vaswani *et al*. [11] achieved huge success in the fields of Natural Language Processing (NLP) and Machine Translation over the last couple of years due to their ability of handling sequential input data by using self-attention [1, 3, 8].

In this supplementary section, we briefly present existing work around the topic of Transformers in the field of medical image segmentation.

1.1. Transformer for medical image segmentation

When it comes to medical image segmentation, the most common architecture used is the U-Net [9] or different variations like U-Net++ [14], no new U-Net (nnU-Net) [6] or Deep Residual U-Net introduced by Zhang *et al*. [13]. It is only recently that architectures for medical image segmentation relying solely on Transformer architectures or hybrid approaches have been presented.

Karimi *et al*. [7] introduce a medical image segmentation network using Transformers instead of a Convolutional Neural Network (CNN). The presented architecture however is very similar to the introduced Vision Transformer (ViT) [11], except it is developed for three-dimensional data like Computer Tomography (CT) scans. The authors show that the positional embedding makes a significant difference in terms of segmentation accuracy.

Another application of Transformer architectures – *TransBTS* – for medical image segmentation is proposed by Wang *et al*. [12] with the specific use for multimodal brain tumor segmentation in Magnetic Resonance Images (MRIs). The presented approach is based on a three-dimensional CNN encoder – decoder combined with a Transformer encoder in between. The Transformer encoder has the same architecture as the ViT.

The *MedT* introduced by Valanarasu *et al*. [10] is a Transformer-based approach for medical image segmentation by using a gated position-sensitive axial attention mechanism consisting of four gates. The authors' evaluation shows that the proposed MedT method outperforms baseline architectures like the U-Net, U-Net++ and Deep Residual U-Net.

1.2. Hybrid U-Net Transformer architectures for medical image segmentation

TransUNet, presented by Chen *et al*. [2], is a hybrid network that combines a CNN-Transformer hybrid model with the conventional U-Net architecture. The authors use CNN in order to extract features and to create a feature map. Regarding the input of the Transformer encoder, patches are extracted from the CNN feature maps instead of the raw input images to increase the performance of the architecture [2]. The encoder is followed with a cascaded upsampler to predict the final segmentation by using multiple upsampling steps.

The proposed *UNETR* [5] follows very closely the structure of the ViT architecture [4]. Three-dimensional volumes are split into three-dimensional patches that are linearly projected and flattened. Those patches are then fed into the Transformer network, whereas different encoded representations from Transformer layers are combined with the decoder using skip connections for predicting the segmentation mask. All in all, the U-Net encoder is replaced with the Transformer encoder and connected to the upsampling decoder that is then used to predict the final segmentation. The authors' evaluation has shown that the presented method outperforms the TransUNet, TransBTS, but also baseline architectures like the nnU-Net.

2. nnU-Net and ViT U-Net comparison in a non-CL setup

In this supplementary part we provide the evaluation results of plain nnU-Nets and ViT U-Nets in a non-CL setup to compare their performances. For this purpose, we train and evaluate for every dataset from the hippocampus corpus one nnU-Net and one ViT U-Net respectively while evaluating the final network on all three hippocampus datasets. Table 1 shows the results based on the Dice scores. Bold values indicate the highest score between the nnU-Net and ViT U-Net. The same evaluation and experimental setups as explained in the main manuscript apply here as well.

Trained on	Architecture	Dice ↑ $\pm\sigma$ ↓ [%]		
		HarP	Dryad	DecathHip
HarP	nnU-Net	85.72 ± 0.77	$\mathbf{84.96 \pm 0.22}$	1.27 ± 0.24
	ViT U-Net	$\mathbf{85.74 \pm 0.99}$	84.81 ± 0.24	$\mathbf{1.59 \pm 0.38}$
Dryad	nnU-Net	$\mathbf{38.76 \pm 5.26}$	90.82 ± 0.27	7.18 ± 1.54
	ViT U-Net	34.63 ± 7.10	$\mathbf{90.96 \pm 0.48}$	$\mathbf{8.01 \pm 2.52}$
DecathHip	nnU-Net	$\mathbf{3.69 \pm 1.20}$	18.31 ± 1.65	89.67 ± 0.40
	ViT U-Net	3.67 ± 1.51	$\mathbf{20.13 \pm 0.98}$	$\mathbf{89.69 \pm 0.39}$

Table 1: nnU-Net and ViT U-Net comparison in terms of performance results based on Dice scores.

Briefly analysing Table 1 it is easy to see that the ViT U-Net architecture outperforms the nnU-Net in two out of three times for every trained dataset. As the ViT U-Net architecture is not the main focus of the manuscript, we do not further analyse this comparison. However, it is worth mentioning that the performance differences are not significant enough to decide which architecture performs best in a non-CL setup.

References

[1] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 1

[2] 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 preprint arXiv:2102.04306*, 2021. 1

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 1

[5] Ali Hatamizadeh, Dong Yang, Holger Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. *arXiv preprint arXiv:2103.10504*, 2021. 1

[6] Fabian Isensee, Jens Petersen, Andre Klein, David Zimmerer, Paul F Jaeger, Simon Kohl, Jakob Wasserthal, Gregor Koehler, Tobias Norajitra, Sebastian Wirkert, et al. nnu-net: Self-adapting framework for u-net-based medical image segmentation. *arXiv preprint arXiv:1809.10486*, 2018. 1

[7] Davood Karimi, Serge Vasylechko, and Ali Gholipour. Convolution-free medical image segmentation using transformers. *arXiv preprint arXiv:2102.13645*, 2021. 1

[8] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *arXiv preprint arXiv:2101.01169*, 2021. 1

[9] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 1

[10] Jeya Maria Jose Valanarasu, Poojan Oza, Ilker Hacihaliloglu, and Vishal M Patel. Medical transformer: Gated axial-attention for medical image segmentation. *arXiv preprint arXiv:2102.10662*, 2021. 1

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 1

[12] Wenxuan Wang, Chen Chen, Meng Ding, Hong Yu, Sen Zha, and Jiangyun Li. Transbts: Multimodal brain tumor segmentation using transformer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 109–119. Springer, 2021. 1

[13] Zhengxin Zhang, Qingjie Liu, and Yunhong Wang. Road extraction by deep residual u-net. *IEEE Geoscience and Remote Sensing Letters*, 15(5):749–753, 2018. 1

[14] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 3–11. Springer, 2018. 1