Can SegFormer be a True Competitor to U-Net for Medical Image Segmentation?

Théo Sourget[1], Syed Nouman Hasany[1], Fabrice Mériaudeau[2], and Caroline Petitjean[1]

[1] Univ Rouen Normandie, Université Le Havre Normandie, INSA Rouen Normandie, Normandie Univ, LITIS UR 4108, 76000 Rouen, France
[2] ICMUB UMR CNRS 6302, Université Bourgogne, 21000 Dijon, France
caroline.petitjean@univ-rouen.fr

Abstract. The U-Net model, introduced in 2015, is established as the state-of-the-art architecture for medical image segmentation, along with its variants UNet++, nnU-Net, V-Net, etc. Vision transformers made a breakthrough in the computer vision world in 2021. Since then, many transformer based architectures or hybrid architectures (combining convolutional blocks and transformer blocks) have been proposed for image segmentation, that are challenging the predominance of U-Net. In this paper, we ask the question whether transformers could overtake U-Net for medical image segmentation. We compare SegFormer, one of the most popular transformer architectures for segmentation, to U-Net using three publicly available medical image datasets that include various modalities and organs: segmentation of cardiac structures in ultrasound images from the CAMUS challenge, segmentation of polyp in endoscopy images and segmentation of instrument in colonoscopy images from the MedAI challenge. We compare them in the light of various metrics (segmentation performance, training time) and show that SegFormer can be a true competitor to U-Net and should be carefully considered for future tasks in medical image segmentation.

Keywords: Medical image segmentation · UNet · transformers

1 Introduction

Since 2015, the U-Net [10] has been established as the state of the art model for medical image segmentation, along with its variants UNet++, nnU-Net [6], V-Net, etc. Its predominance has been challenged by the arrival of transformers in 2021 [2]. Indeed following the excellent results of Transformers on natural language processing problems, transformer-based architectures have been proposed on image processing tasks, starting with the Vision Transformer, for image classification [3]. Transformers process the image as a sequence of patches (typically of size 16×16 pixels) and seem to be very efficient thanks to the attention mechanism which allows them to capture long range interaction between the patches - contrary to the reduced receptive field of convolutional kernels.

The transition from the ViT to an architecture with transformers for image segmentation is not obvious. Many architectures have been proposed, such as the Segmenter Transformer SETR [13] or the PVT [11] or SegFormer, whose transformer encoders are based on a pyramid structure, so as to mimic the encoder of a CNN. The recently released "Segment Anything" tool [7], a promptable segmentation system with impressive performance on natural images, is also based on a transformer architecture. Hybrid architectures combining convolutional and transformer blocks such as TransUNet [2], CATS [9] or UNETR [4] have been proposed for medical image segmentation, as underlined by recent reviews of transformers in medical image analysis [1,5]. These models are however often very complex with several tens of millions of parameters and require lot of time to be trained.

The question now is whether a transformer-based architecture can be a true competitor to U-Net, for medical image segmentation. In this study, we decide to focus on SegFormer [12], a lightweight architecture for image segmentation, designed to avoid complex decoders. Its efficiency, accuracy, and robustness have been shown on a variety of datasets such as ADE20K, Cityscapes, and COCO-Stuff; and in particular, it is shown in the paper to outperform the previous best method on ADE20K, the SETR[13] model . In this paper, our goal is to test SegFormer, both pre-trained and trained from scratch, against the U-Net architecture, in terms of segmentation accuracy and compute efficiency and see if SegFormer can be a viable alternative to U-Net for medical image segmentation.

In the following, we first detail the SegFormer architecture. Then we introduce three datasets encompassing different tasks of binary and multilabel segmentation in medical images, and present the experimental protocol to make a fair comparison. We provide both quantitative and qualitative segmentation results of the models under scrutiny.

2 SegFormer

SegFormer has two main modules: a hierarchically structured transformer encoder and an MLP (multi-layer perceptron) decoder (see Figure 1). One of the key contribution of this architecture is the lightweight All-MLP decoder, resulting in a light architecture in comparison to other transformer architectures for segmentation, e.g. [13]. These aspects of SegFormer lead to multiple benefits. First, as the encoder part produces multiple level feature maps fused in the decoder, the model is able to capture both high and low resolution information. The encoder also relies on a "mix-FFN" (feed-forward network) operation, where a 3×3 convolution with 0-padding and an MLP are mixed into each FFN, to replace the original positional encoding used by other architectures. Moreover, as the model is less complex than other transformer based architectures, it requires less data to be trained and can also be applied to real-time application. Finally, the encoder part of the architecture can be scaled from B0 to B5 by increasing the number of layers or the dimensions of encoder blocks ; depending on the need, it is possible to favor time efficiency with B0 able to perform

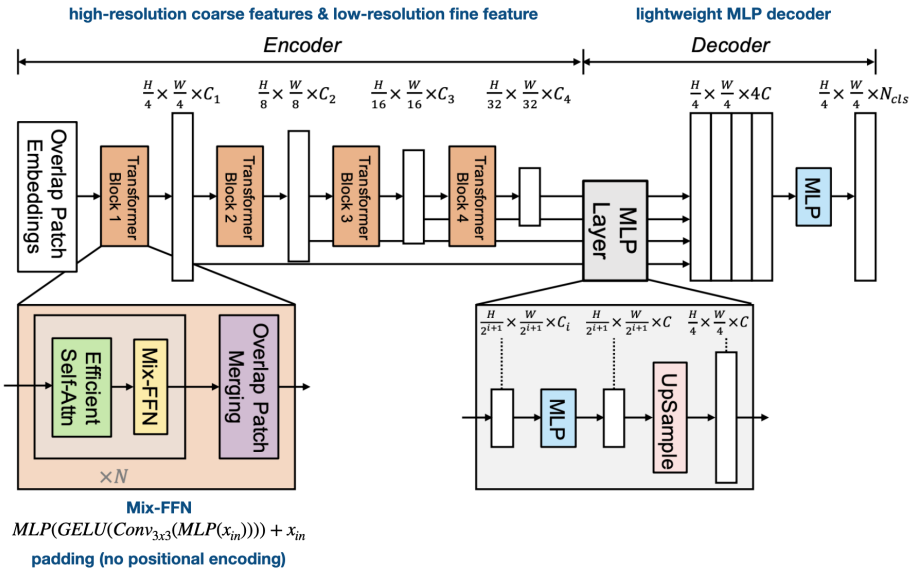


Fig. 1: SegFormer architecture. The two main modules are the hierarchical Transformer encoder, and an all-MLP based decoder merging the features from the encoder. Figure reproduced from [12].

real-time applications or performances with B5 obtaining best results on various problems. In this study, we will use the SegFormer-B0 architecture that has 3.1M parameters.

3 Experimentations

3.1 Datasets

We have assessed the two architectures on three different datasets: CAMUS, Polyp, and Instruments. The Cardiac Acquisitions for Multi-structure Ultrasound Segmentation (CAMUS) challenge dataset[8] contains 2D images of cardiac ultrasound images from 500 patients, with the manual segmentation of 4 different classes (Endocardium, Epicardium, Atrium and Background). For each patient, there are 4 different types of images related to different views (2-chambers or 4-chambers) and phase of the heart (diastole and systole). In this experiment, we have used only the diastole images for both training and testing phase. Poly and Instruments are two datasets from the MedAI: Transparency in Medical Image Segmentation challenge. The Polyp Segmentation Task consists in segmenting polyps in endoscopy images from the Kvasir-SEG dataset which contains 1000 images and corresponding binary segmentation masks. The Instrument Segmentation Task is about segmenting instruments present in colonoscopy

Table 1: Dataset size

Dataset size	Train	Valid	Test
Camus	500	450	50
Polyp	640	160	200
Instrument	377	95	118

videos from the Kvasir-Instrument dataset which contains 590 images and corresponding binary segmentation masks.

3.2 Protocol

In our experiments, we compare the original U-Net (31M parameters) to 2 versions of the SegFormer-B0 model (3.7M parameters): one pre-trained on ImageNet-1k and the other trained from scratch. We also introduce a U-Net Lite, a lightweight version of U-Net where we have reduced the number of filters per layer from [64,128,256,512,1024] in the original paper [10] to [22,44,88,176,352] to match the number of parameters of SegFormer-B0, i.e. 3.7M. For the CAMUS dataset the loss function is Cross-Entropy, and the optimizer is Adam with a fixed learning rate, 1e-03 for U-Net, and 1e-04 for SegFormer. For the Polyp and Instrument segmentation tasks, the loss function is an average of Cross-Entropy and Dice, and the optimizer is AdamW with a fixed learning rate of 1e-04 for both the U-Net and the SegFormer.

We use the SegFormer-B0 model from HuggingFace and we applied transfer learning by using encoder's weights trained on Imagenet-1k. For sake of reproducibility, the code used on CAMUS for U-Net can be found here: https://github.com/TheoSourget/UNet_CAM and for SegFormer here: https://github.com/TheoSourget/SegFormer_CAMUS.

Table 1 shows the datasets split into training, validation and test sets. For CAMUS, the images are resized to 256×256 pixels and at every epoch, a random rotation between -10° and 10° is applied to perform data augmentation on the dataset. For Polyp and Instrument Segmentation, the images are resized to 224×224 and random flipping (horizontal and vertical) as well as random rotation between 0° and 180° is applied to perform data augmentation on the dataset.

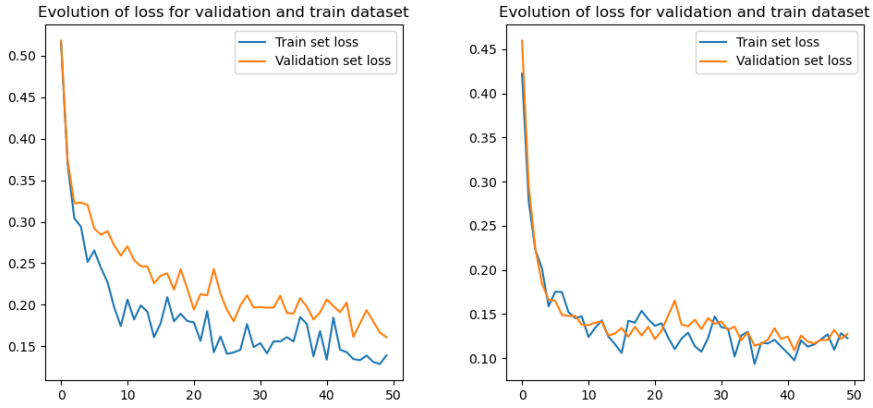
4 Results and Discussion

4.1 Segmentation accuracy

The average Dice Scores for each dataset are in Table 2 and Figure 4. First of all, it is interesting to note that even if the number of parameters of U-Net is drastically reduced, the decrease in accuracy on the CAMUS dataset is not as significant than for Polyp or Instrument. This corroborates the observations made in [**?**,8] that simpler models can obtain similar results than more complex ones. Not surprisingly, the pre-trained SegFormer is better, sometimes by a large margin, than the SegFormer trained from scratch. This is also visible by the the

Table 2: Results: Average Dice scores of U-Net and SegFormer of 3 datasets: CAMUS, Polyp and Instrument. * indicates that the score is significantly different from that of UNet ($p<0.05$).For graphical representation see Figure 4

		U-Net	U-Net Lite	SegFormer	SegFormer pre-trained
Pre-trained?		No	No	No	Yes
# param		31M	3.7M	3.7M	3.7M
CAMUS	Endo	0.90	0.90	0.89	**0.91***
	Epi	0.80	0.79	0.81	**0.83***
	Atrium	0.83	0.84	0.81	0.85
Polyp		0.74	0.67	0.60	**0.83***
Instrum		0.79	0.75	0.82	**0.92***



(a) SegFormer (trained from scratch) (b) pre-trained SegFormer

Fig. 2: Evolution of loss during training for SegFormer with and without pre-training on ImageNet-1K

behaviour during training: the pre-trained model seems to converge faster than the one trained from scratch, as shown by the evolution of loss for the CAMUS dataset in Fig 2.

Finally, the pre-trained SegFormer performed significantly better than U-Net for almost all segmented regions, i.e. the endocardium and epicardium of CAMUS, Polyp and Instrument, except for the atrium of CAMUS. Statistical significance was achieved using a two-sided Wilcoxon test ($p<0.05$). We can also see a difference in the training behaviour: while the transformer is able to predict every class at the end of the first epoch, U-Net only predicts the background class during the first two epochs and only afterwards is able to predict relevant classes. This is illustrated in Figure 3 where the difference of segmen-

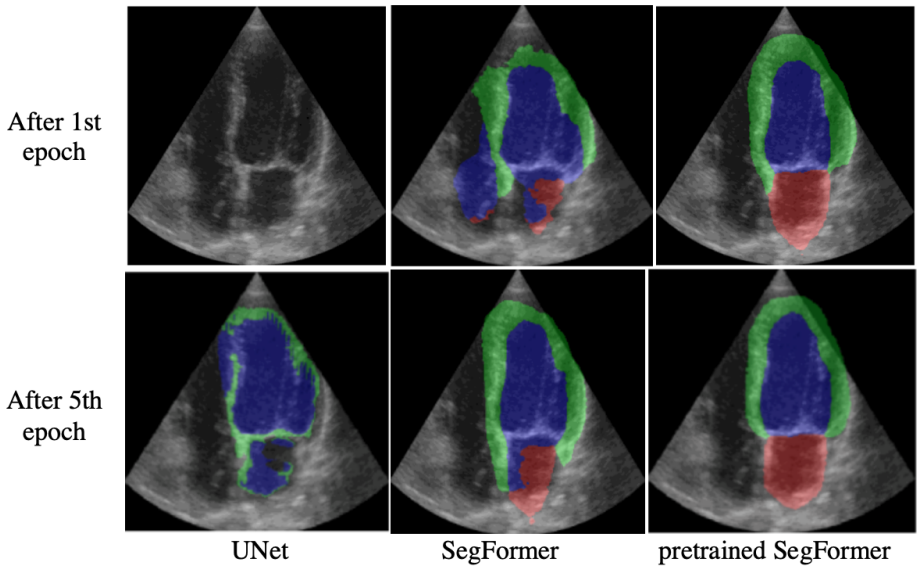


Fig. 3: Comparing segmentation results between U-Net, SegFormer and pre-trained SegFormer after the 1st and 5th epoch

Table 3: Drop in training time of U-Net Lite, SegFormer and pre-trained SegFormer with respect to U-Net's training time.

Model	U-Net Lite	SegFormer	SegFormer pre-trained	Epochs
CAMUS	-57.5%	-49.7%	-53.0%	50
Polyp	-51.2%	-62.3%	-65.1%	80
Instrum	-40.4%	-46.4%	-46.9%	80

tation between U-Net, SegFormer and the pre-trained SegFormer after 1st and 5th epoch on a validation example from CAMUS dataset can be seen. This Figure also highlights the strong impact of using pre-trained weights for SegFormer.

4.2 Training time

In Table 3, we display the decrease in training time over the indicated number of epochs of the considered models, with respect to that of U-Net. We can gather from this table that SegFormer is always faster to train than the original U-Net and often faster than the Lite U-Net version, even if it is not pre-trained.

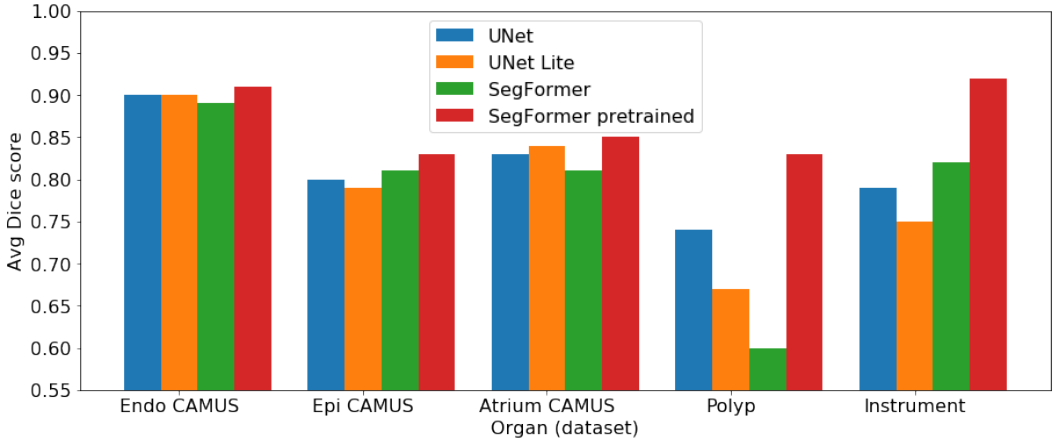


Fig. 4: Average Dice scores of U-Net and SegFormer on test sets of 3 datasets: CAMUS, Polyp and Instrument. Corresponds to results in Table 2.

5 Conclusion

In this study, our goal was to compare the U-Net model to a newly proposed, transformer based model called SegFormer. The underlying aim was to start assessing whether the breakthrough of vision transformer can really offer competitive segmentation performance both in accuracy and training time, on some medical image segmentation tasks, namely here the segmentation of cardiac structured in ultrasound images, of polyp in endoscopy, and of instruments in colonoscopy. On every task, pre-trained SegFormer-B0 obtained on par or better results than U-Net, and in less training time than the original U-Net thanks to its light architecture. Hence, we have shown that even if transformers usually need more data to be trained, they can still be applied on medical imaging tasks and obtained better results with limited dataset, taking advantage of transfer learning. However, this is only a preliminary study, and we are working towards large scale experiments, to include other datasets, to assess whether this findings holds on more segmentation tasks.

Acknowledgments

The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR), under grant Project-ANR-21-CE23-0013 (project MediSEG).

References

1. Azad, R., Kazerouni, A., Heidari, M., Aghdam, E.K., Molaei, A., Jia, Y., Jose, A., Roy, R., Merhof, D.: Advances in Medical Image Analysis with Vision Transformers: A Comprehensive Review (Jan 2023).

https://doi.org/10.48550/arXiv.2301.03505, http://arxiv.org/abs/2301.03505, arXiv:2301.03505 [cs]
2. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, L.A., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation (2021)
3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. CoRR **abs/2010.11929** (2020), https://arxiv.org/abs/2010.11929
4. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, A.B., Roth, R.H., Xu, D.: Unetr - transformers for 3d medical image segmentation. WACV pp. 1748–1758 (2022)
5. He, K., Gan, C., Li, Z., Rekik, I., Yin, Z., Ji, W., Gao, Y., Wang, Q., Zhang, J., Shen, D.: Transformers in medical image analysis. Intelligent Medicine **3**(1), 59–78 (Feb 2023). https://doi.org/10.1016/j.imed.2022.07.002, https://www.sciencedirect.com/science/article/pii/S2667102622000717
6. Isensee, F., Petersen, J., Klein, A., Zimmerer, D., Jaeger, P.F., Kohl, S., Wasserthal, J., Köhler, G., Norajitra, T., Wirkert, S.J., Maier-Hein, K.H.: nnu-net: Self-adapting framework for u-net-based medical image segmentation. CoRR **abs/1809.10486** (2018), http://arxiv.org/abs/1809.10486
7. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything (2023)
8. Leclerc, S., Smistad, E., Pedrosa, J., Ostvik, A., Cervenansky, F., Espinosa, F., Espeland, T., Berg, E.A.R., Jodoin, P.M., Grenier, T., Lartizien, C., Dhooge, J., Lovstakken, L., Bernard, O.: Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. IEEE Transactions on Medical Imaging **38**(9), 2198–2210 (sep 2019). https://doi.org/10.1109/tmi.2019.2900516, https://doi.org/10.1109%2Ftmi.2019.2900516
9. Li, H., Hu, D., Liu, H., Wang, J., Oguz, I.: Cats: Complementary cnn and transformer encoders for segmentation. In: 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI). pp. 1–5 (2022). https://doi.org/10.1109/ISBI52829.2022.9761596
10. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. MICCAI p. 234–241 (2015)
11. Wang, W., Xie, E., Li, X., Fan, D., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. CoRR **abs/2102.12122** (2021), https://arxiv.org/abs/2102.12122
12. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. CoRR **abs/2105.15203** (2021), https://arxiv.org/abs/2105.15203
13. Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H.S., Zhang, L.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. CoRR **abs/2012.15840** (2020), https://arxiv.org/abs/2012.15840