U-Net Transformer: Self and Cross Attention for Medical Image Segmentation

Olivier Petit[1,2], Nicolas Thome[1], Clement Rambour[1], Loic Themyr[1,3], Toby Collins[3], and Luc Soler[2]

[1] CEDRIC - Conservatoire National des Arts et Metiers, Paris, France
[2] Visible Patient SAS, Strasbourg, France
[3] IRCAD, Strasbourg, France
olivier.petit@visiblepatient.com

Abstract. Medical image segmentation remains particularly challenging for complex and low-contrast anatomical structures. In this paper, we introduce the U-Transformer network, which combines a U-shaped architecture for image segmentation with self- and cross-attention from Transformers. U-Transformer overcomes the inability of U-Nets to model long-range contextual interactions and spatial dependencies, which are arguably crucial for accurate segmentation in challenging contexts. To this end, attention mechanisms are incorporated at two main levels: a self-attention module leverages global interactions between encoder features, while cross-attention in the skip connections allows a fine spatial recovery in the U-Net decoder by filtering out non-semantic features. Experiments on two abdominal CT-image datasets show the large performance gain brought out by U-Transformer compared to U-Net and local Attention U-Nets. We also highlight the importance of using both self- and cross-attention, and the nice interpretability features brought out by U-Transformer.

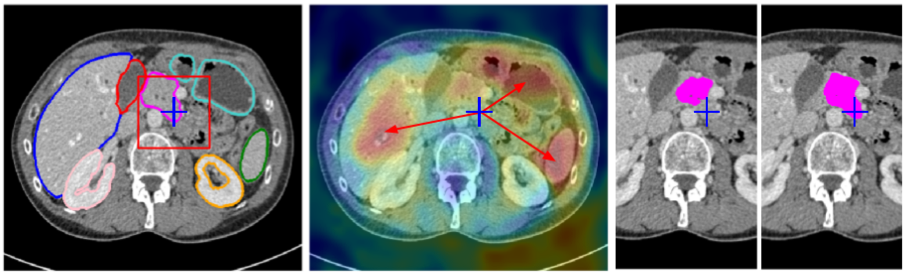
Keywords: Medical Image Segmentation · Transformers · Self-attention · Cross-attention · Spatial layout · Global interactions

1 Introduction

Organ segmentation is of crucial importance in medical imaging and computed-aided diagnosis, *e.g.* for radiologists to assess physical changes in response to a treatment or for computer-assisted interventions.

Currently, state-of-the-art methods rely on Fully Convolutional Networks (FCNs), such as U-Net and variants [9, 2, 7, 18]. U-Nets use an encoder-decoder architecture: the encoder extracts high-level semantic representations by using a cascade of convolutional layers, while the decoder leverages skip connections to re-use high-resolution feature maps from the encoder in order to recover lost spatial information from high-level representations.

Despite their outstanding performances, FCNs suffer from conceptual limitations in complex segmentation tasks, *e.g.* when dealing with local visual ambiguities and low contrast between organs. This is illustrated in Fig 1a) for



a) Ground Truth b) Attention map c) U-Net d) U-Transformer

Fig. 1. Global context is crucial for complex organ segmentation but cannot be captured by vanilla U-Nets with a limited receptive field, *i.e.* blue cross region in a) with failed segmentation in c). The proposed U-Transformer network represents full image context by means of attention maps b), which leverage long-range interactions with other anatomical structures to properly segment the complex pancreas region in d).

segmenting the blue cross region corresponding to the pancreas with U-Net: the limited Receptive Field (RF) framed in red does not capture sufficient contextual information, making the segmentation fail, see Fig 1c).

In this paper, we introduce the U-Transformer network, which leverages the strong abilities of transformers [13] to model long-range interactions and spatial relationships between anatomical structures. U-Transformer keeps the inductive bias of convolution by using a U-shaped architecture, but introduces attention mechanisms at two main levels, which help to interpret the model decision. Firstly, a self-attention module leverages global interactions between semantic features at the end of the encoder to explicitly model full contextual information. Secondly, we introduce cross-attention in the skip connections to filter out non-semantic features, allowing a fine spatial recovery in the U-Net decoder.

Fig 1b) shows a cross-attention map induced by U-Transformer, which highlights the most important regions for segmenting the blue cross region in Fig 1a): our model leverages the long-range interactions with respect to other organs (liver, stomach, spleen) and their positions to properly segment the whole pancreas region, see Fig 1d). Quantitative experiments conducted on two abdominal CT-image datasets show the large performance gain brought out by U-Transformer compared to U-Net and to the local attention in [11].

Related Work. Attention mechanisms are a relatively recent problem in medical imaging [16, 8, 10–12]. Attention in segmentation is often based on multi-resolution features combined with a simple attention module [16, 6]. These contributions however fail to incorporate long-range dependencies. Recent works successfully tackle this aspect through Dual attention networks [12, 5] proving the importance of full range attention but to the cost of large parameter overhead and multiple concurrent loss functions.

Transformers [13] models also bring global attention and have witnessed increasing success in the last five years, started in natural language processing with text embeddings [3]. A pioneer use of transformers in computer vision is non-local networks [15], which combine self-attention with a convolutional backbone. Recent applications include object detection [1], semantic segmentation [17, 14], and image classification [4].

U-Transformer combines the power of Transformers to grasp long-range dependencies and multi-resolution information processing through self- and cross-attention modules. Our cross-attention mechanism shares the high-level motivation of Attention U-Net [11] to help the recovery of fine spatial information from rich semantic features, with the noticeable difference that the U-Transformer's attention embraces all input features whereas Attention U-Net's attention uses each local feature independently.

2 The U-Transformer Network

As mentioned in Section 1, encoder-decoder U-shaped architectures lack global context information to handle complex medical image segmentation tasks. We introduce the U-Transformer network, which augments U-Nets with attention modules built from multi-head transformers. U-Transformer models long-range contextual interactions and spatial dependencies by using two types of attention modules (see Fig 2): Multi-Head Self-Attention (MHSA) and Multi-Head Cross-Attention (MHCA). Both modules are designed to express a new representation of the input based on its self-attention in the first case (*cf.* 2.1) or on the attention paid to higher level features in the second (*cf.* 2.2).

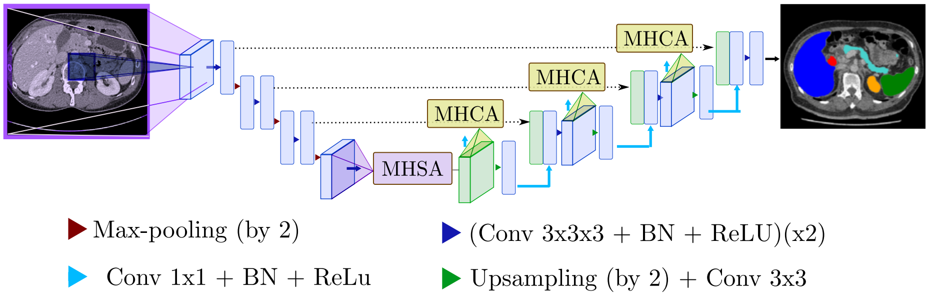

MHCA
MHCA
MHCA
MHSA
▶ Max-pooling (by 2)
▶ (Conv 3x3x3 + BN + ReLU)(x2)
▶ Conv 1x1 + BN + ReLu
▶ Upsampling (by 2) + Conv 3x3

Fig. 2. U-Transformer augments U-Nets with transformers to model long-range contextual interactions. The Multi-Head Self-Attention (MHSA) module at the end of the U-Net encoder gives access to a receptive field containing the whole image (shown in purple), in contrast to the limited U-Net receptive field (shown in blue). Multi-Head Cross-Attention (MHCA) modules are dedicated to combine the semantic richness in high level feature maps with the high resolution ones coming from the skip connections.

2.1 Self-attention

The MHSA module is designed to extract long range structural information from the images. To this end, it is composed of multi-head self-attention functions as described in [13] positioned at the bottom of the U-Net as shown in Figure 2.

The main goal of MHSA is to connect every element in the highest feature map with each other, thus giving access to a receptive field including all the input image. The decision for one specific pixel can thus be influenced by any input pixel. The attention formulation is given in Equation 1. A self-attention module takes three inputs, a matrix of queries $\boldsymbol{Q} \in \mathbb{R}^{n \times d_k}$, a matrix of keys $\boldsymbol{K} \in \mathbb{R}^{n \times d_k}$ and a matrix of values $\boldsymbol{V} \in \mathbb{R}^{n \times d_k}$.

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}})\boldsymbol{V} = \boldsymbol{A}\boldsymbol{V} \tag{1}$$

A line of the attention matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ corresponds to the similarity of a given element in $\boldsymbol{Q}$ with respect to all the elements in $\boldsymbol{K}$. Then, the attention function performs a weighted average of the elements of the value $\boldsymbol{V}$ to account for all the interactions between the queries and the keys as illustrated in Figure 3. In our segmentation task, $\boldsymbol{Q}$, $\boldsymbol{K}$ and $\boldsymbol{V}$ share the same size and correspond to different learnt embedding of the highest level feature map denoted by $\boldsymbol{X}$ in Figure 3. The embedding matrices are denoted as $\boldsymbol{W}_q$, $\boldsymbol{W}_k$ and $\boldsymbol{W}_v$. The attention is calculated separately in multiple heads before being combined through another embedding. Moreover, to account for absolute contextual information, a positional encoding is added to the input features. It is especially relevant for medical image segmentation, where the different anatomical structures follow a fixed spatial position. The positional encoding can thus be leveraged to capture absolute and relative position between organs in MHSA.

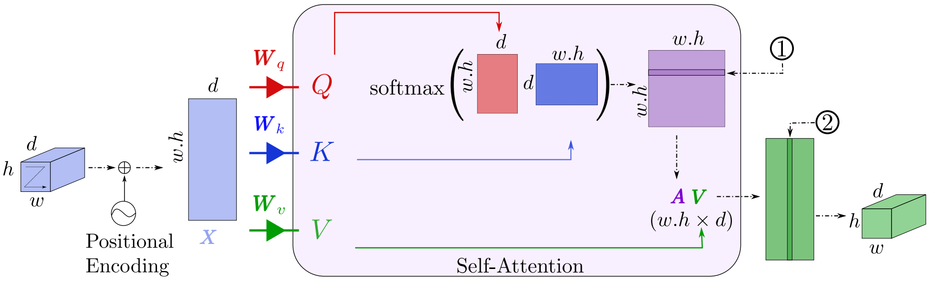


Fig. 3. MHSA module: the input tensor is embedded into a matrix of queries $\boldsymbol{Q}$, keys $\boldsymbol{K}$ and values $\boldsymbol{V}$. The attention matrix $\boldsymbol{A}$ in purple is computed based on $\boldsymbol{Q}$ and $\boldsymbol{K}$. (1) A line of $\boldsymbol{A}$ corresponds to the attention given to all the elements in $\boldsymbol{K}$ with respect to one element in $\boldsymbol{Q}$. (2) A column of the value $\boldsymbol{V}$ corresponds to a feature map weighted by the attention in $\boldsymbol{A}$.

2.2 Cross-attention

The MHSA module allows to connect every element in the input with each other. Attention may also be used to increase the U-Net decoder efficiency and in particular enhance the lower level feature maps that are passed through the skip connections. Indeed, if these skip connections insure to keep a high resolution information they lack the semantic richness that can be found deeper in the network. The idea behind the MHCA module is to turn off irrelevant or noisy areas from the skip connection features and highlight regions that present a significant interest for the application. Figure 4 shows the cross-attention module. The MHCA block is designed as a gating operation of the skip connection $\boldsymbol{S}$ based on the attention given to a high level feature map $\boldsymbol{Y}$. The computed weight values are then re-scaled between 0 and 1 through a sigmoid activation function. The resulting tensor, denoted $\boldsymbol{Z}$ in Figure 4, is a filter where low magnitude elements indicate noisy or irrelevant areas to be reduced. A cleaned up version of $\boldsymbol{S}$ is then given by the Hadamard product $\boldsymbol{Z} \odot \boldsymbol{S}$. Finally, the result of this filtering operation is concatenated with the high level feature tensor $\boldsymbol{Y}$. Here, the keys and queries are computed from the same source as we are designing a filtering operation whereas for NLP tasks, having homogeneous keys and values may be more meaningful. This configuration proved to be empirically more effective.

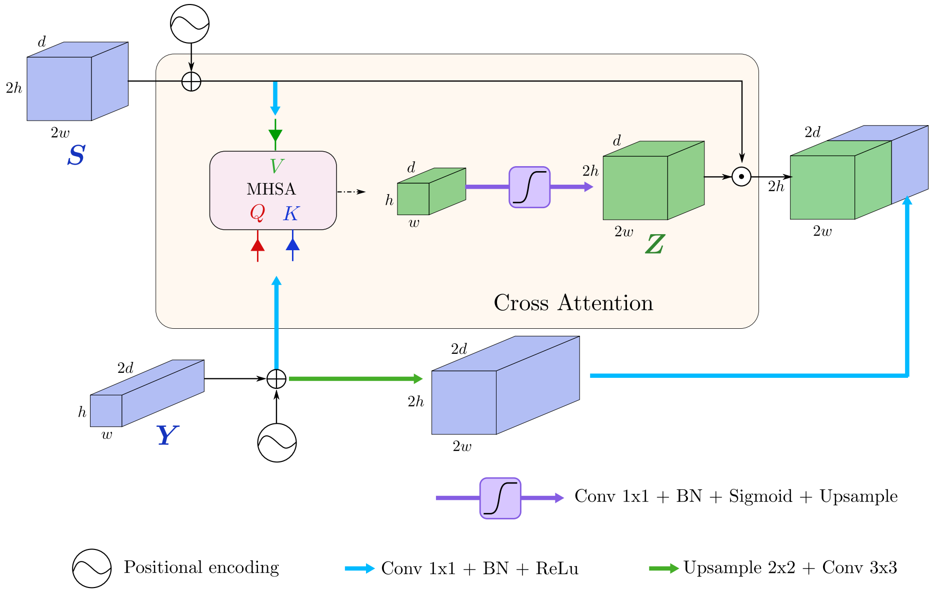


Fig. 4. MHCA module: the value of the attention function corresponds to the skip connection $\boldsymbol{S}$ weighted by the attention given to the high level feature map $\boldsymbol{Y}$. This output is transformed into a filter $\boldsymbol{Z}$ and applied to the skip connection.

3 Experiments

We evaluate U-Transformer for abdominal organ segmentation on the TCIA pancreas public dataset, and an internal multi-organ dataset.

Accurate pancreas segmentation is particularly difficult, due to its small size, complex and variable shape, and because of the low contrast with the neighboring structures, see Fig 1. In addition, the multi-organ setting assesses how U-transformer can leverage attention from multi-organ annotations.

Experimental setup The TCIA pancreas dataset[4] contains 82 CT-scans with pixel-level annotations. Each CT-scan has around $181 \sim 466$ slices of 512×512 pixels and a voxel spacing of $([0.66 \sim 0.98] \times [0.66 \sim 0.98] \times [0.5 \sim 1.0])$ mm^3.

We also experiment with an Internal Multi-Organ (IMO) dataset composed of 85 CT-scans annotated with 7 classes: liver, gallbladder, pancreas, spleen, right and left kidneys, and stomach. Each CT-scan has around $57 \sim 500$ slices of 512×512 pixels and a voxel spacing of $([0.42 \sim 0.98] \times [0.42 \sim 0.98] \times [0.63 \sim 4.00])$mm^3.

All experiments follow a 5-fold cross validation, using 80% of images in training and 20% in test. We use the Tensorflow library to train the model, with Adam optimizer (10^{-4} learning rate, exponential decay scheduler).

We compare U-Transformer to the U-Net baseline [9] and Attention U-Net [11] with the same convolutional backbone for fair comparison. We also report performances with self-attention only (MHSA, section 2.1), and the cross-attention only (MHCA, section 2.2). U-Net has ~ 30M parameters, the overhead from U-transformer is limited (MHSA ~ 5M, each MHCA block ~ 2.5M).

3.1 U-Transformer performances

Table 1 reports the performances in Dice averaged over the 5 folds, and over organs for IMO. U-Transformer outperforms U-Net by 2.4pts on TCIA and 1.3pts for IMO, and Attention U-Net by 1.7pts for TCIA and 1.6pts for IMO. The gains are consistent on all folds, and paired t-tests show that the improvement is significant with $p-$values $< 3\%$ for every experiment.

Table 1. Results for each method in Dice similarity coefficient (DSC, %)

Dataset	U-Net [9]	Attn U-Net [11]	MHSA	MHCA	U-Transformer
TCIA	76.13 (± 0.94)	76.82 (± 1.26)	77.71 (± 1.31)	77.84 (± 2.59)	**78.50** (± 1.92)
IMO	86.78 (± 1.72)	86.45 (± 1.69)	87.29 (± 1.34)	87.38 (± 1.53)	**88.08** (± 1.37)

Figure 5 provides qualitative segmentation comparison between U-Net, Attention U-Net and U-Transformer. We observe that U-Transformer performs better on difficult cases, where the local structures are ambiguous. For example, in the second row, the pancreas has a complex shape which is missed by U-Net and Attention U-Net but U-Transformer successfully segments the organ.

In Table 1, we can see that the self-attention (MHSA) and cross-attention (MHCA) alone already outperform U-Net and Attention U-Net on TCIA and IMO. Since MHCA and Attention U-Net apply attention mechanisms at the skip connection level, it highlights the superiority of modeling global interactions between anatomical structures and positional information instead of the simple

[4] https://wiki.cancerimagingarchive.net/display/Public/Pancreas-CT

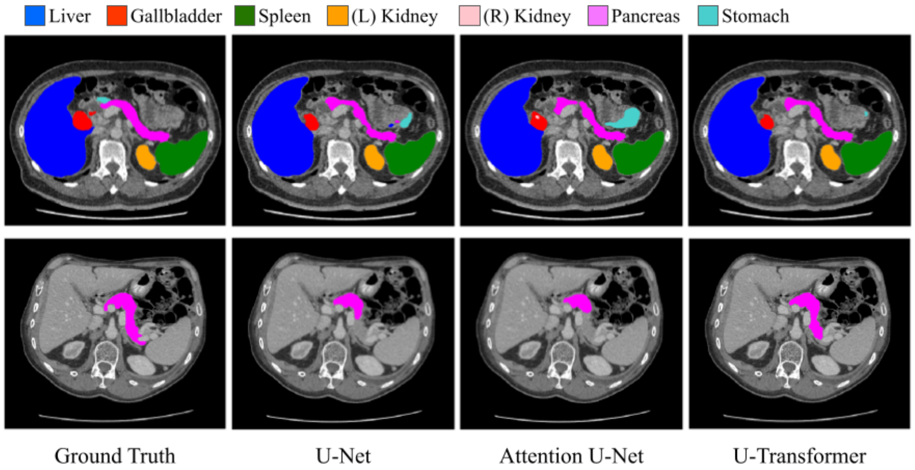


Fig. 5. Segmentation results for U-Net [9], Attention U-Net [11] and U-Transformer on the multi-organ IMO dataset (first row) and on TCIA pancreas (second row).

local attention in [11]. Finally, the combination of MHSA and MHCA in U-Transformer shows that the two attention mechanisms are complementary and can collaborate to provide better segmentation predictions.

Table 2 details the results for each organ on the multi-organ IMO dataset. This further highlights the interest of U-Transformer, which significantly outperforms U-Net and Attention U-Net for the most challenging organs: pancreas: +3.4pts, gallbladder: +1.3pts and stomach: +2.2pts. This validates the capacity of U-Transformer to leverage multi-label annotations to drive the interactions between anatomical structures, and use easy organ predictions to improve the detection and delineation of more difficult ones. We can note that U-Transformer is better for every organ, even the liver which has a high score $> 95\%$ with U-Net.

Table 2. Results on IMO in Dice similarity coefficient (DSC, %) detailed per organ.

Organ	U-Net [9]	Attn U-Net [11]	MHSA	MHCA	U-Transformer
Pancreas	69.71 (± 3.74)	68.65 (± 2.95)	71.64 (± 3.01)	71.87 (± 2.97)	**73.10** (± 2.91)
Gallbladder	76.98 (± 6.60)	76.14 (± 6.98)	76.48 (± 6.12)	77.36 (± 6.22)	**78.32** (± 6.12)
Stomach	83.51 (± 4.49)	82.73 (± 4.62)	84.83 (± 3.79)	84.42 (± 4.35)	**85.73** (± 3.99)
Kidney(R)	92.36 (± 0.45)	92.88 (± 1.79)	92.91 (± 1.84)	92.98 (± 1.70)	**93.32** (± 1.74)
Kidney(L)	93.06 (± 1.68)	92.89 (± 0.64)	92.95 (± 1.30)	92.82 (± 1.06)	**93.31** (± 1.08)
Spleen	95.43 (± 1.76)	95.46 (± 1.95)	95.43 (± 2.16)	95.41 (± 2.21)	**95.74** (± 2.07)
Liver	96.40 (± 0.72)	96.41 (± 0.52)	96.82 (± 0.34)	96.79 (± 0.29)	**97.03** (± 0.31)

3.2 U-Transformer analysis and properties

Positional encoding and multi-level MHCA. The Positional Encoding (PE) allows to leverage the absolute position of the objects in the image. Table 3

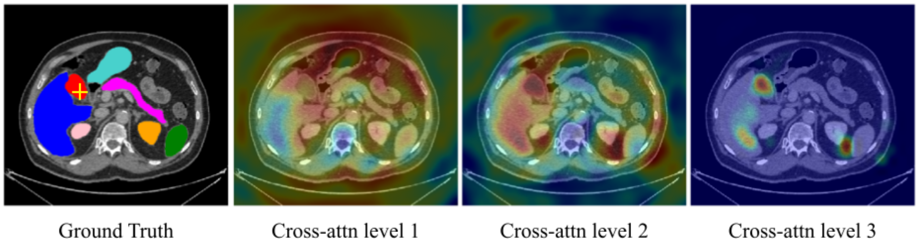


Fig. 6. Cross-attention maps for the yellow-crossed pixel (left image).

shows an analysis of its impact, on one fold on both datasets. For MHSA, the PE improves the results by +0.7pt for TCIA and +0.6pt for IMO. For MHCA, we evaluate a single level of attention with and without PE. We can observe an improvement of +1.7pts for TCIA and +0.6pt for IMO between the two versions.

Table 3 also shows the favorable impact of using multi *vs* single-level attention for MHCA: +1.8pts for TCIA and +0.6pt for IMO. It is worth noting that Attention U-Net uses multi-level attention but remains below MHCA with a single level. Figure 6 shows attention maps at each level of U-Transformer: level 3 corresponds to high-resolution features maps, and tends to focus on more specific regions compared to the first levels.

Table 3. Ablation study on the positional encoding and multi-level on one fold of TCIA and IMO.

			MHSA		MHCA		
	U-Net	Attn U-Net	wo PE –	w PE	1 lvl wo PE –	1 lvl w PE –	multi-lvl w PE
TCIA	76.35	77.23	78.17	**78.90**	77.18	78.88	**80.65**
IMO	88.18	87.52	88.16	**88.76**	87.96	88.52	**89.13**

Further analysis. To further analyse the behaviour of U-Transformer, we evaluate the impact of the number of attention heads for MHSA (supplementary, Fig 1): more heads lead to better performances, but the biggest gain comes from the first head (*i.e.* U-Net to MHSA). Finally, the evaluation of U-Transformer with respect to the Hausdorff distance (supplementary, Table 1) follows the same trend than with Dice score. This highlights the capacity of U-Transformer to reduce prediction artefacts by means of self- and cross-attention.

4 Conclusion

This paper introduces the U-Transformer network, which augments a U-shaped FCN with Transformers. We propose to use self and cross-attention modules to model long-range interactions and spatial dependencies. We highlight the relevance of the approach for abdominal organ segmentation, especially for small and complex organs. Future works could include the study of U-Transformer in 3D networks, with other modalities such as MRI or US images, as well as for other medical image tasks.

References

1. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European Conference on Computer Vision. pp. 213–229. Springer (2020)
2. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: Learning dense volumetric segmentation from sparse annotation. In: MICCAI. pp. 424–432 (2016)
3. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. CoRR **abs/1810.04805** (2018), http://arxiv.org/abs/1810.04805
4. 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. In: International Conference on Learning Representations (2021)
5. Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H.: Dual attention network for scene segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
6. Li, C., Tong, Q., Liao, X., Si, W., Sun, Y., Wang, Q., Heng, P.A.: Attention based hierarchical aggregation network for 3d left atrial segmentation. In: Statistical Atlases and Computational Models of the Heart. Atrial Segmentation and LV Quantification Challenges. pp. 255–264 (2019)
7. Milletari, F., Navab, N., Ahmadi, S.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV). pp. 565–571 (2016)
8. Nie, D., Gao, Y., Wang, L., Shen, D.: Asdnet: Attention based semi-supervised deep networks for medical image segmentation. In: Frangi, A., Fichtinger, G., Schnabel, J., Alberola-López, C., Davatzikos, C. (eds.) MICCAI 2018. pp. 370–378. Lecture Notes in Computer Science, Springer Verlag (2018)
9. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241 (2015)
10. Roy, A.G., Navab, N., Wachinger, C.: Concurrent spatial and channel squeeze & excitation in fully convolutional networks. In: MICCAI. vol. abs/1803.02579 (2018)
11. Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B., Rueckert, D.: Attention gated networks: Learning to leverage salient regions in medical images. Medical Image Analysis **53** (02 2019). https://doi.org/10.1016/j.media.2019.01.012
12. Sinha, A., Dolz, J.: Multi-scale self-guided attention for medical image segmentation. IEEE Journal of Biomedical and Health Informatics pp. 1–1 (2020)
13. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: NeurIPS. pp. 5998–6008 (2017)
14. Wang, H., Zhu, Y., Green, B., Adam, H., Yuille, A., Chen, L.C.: Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In: European Conference on Computer Vision. pp. 108–126 (2020)
15. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7794–7803 (2018)
16. Wang, Y., Deng, Z., Hu, X., Zhu, L., Yang, X., xu, X., Heng, P.A., Ni, D.: Deep attentional features for prostate segmentation in ultrasound. In: MICCAI (09 2018)

17. Ye, L., Rochan, M., Liu, Z., Wang, Y.: Cross-modal self-attention network for referring image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10502–10511 (2019)
18. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. pp. 3–11 (2018)

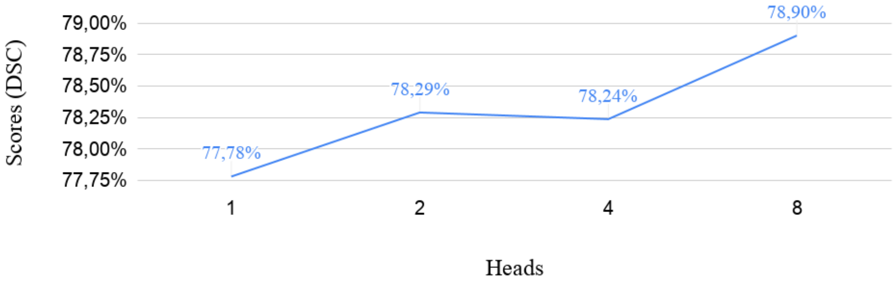


Fig. 7. Evolution of the Dice Score on TCIA (fold 1) when the number of heads varies between 0 and 8 in MHSA.

Table 4. Hausdorff Distances (HD) for the different models

Dataset	U-Net	Attn U-Net	U-Transformer
TCIA	13.61 (± 2.01)	12.48 (± 1.36)	**12.34** (± 1.51)
IMO	12.06 (± 1.65)	12.13 (± 1.58)	**12.00** (± 1.32)