UNETR: Transformers for 3D Medical Image Segmentation

Ali Hatamizadeh, Dong Yang, Holger Roth, and Daguang Xu

NVIDIA, Santa Clara, CA, USA

Abstract. Fully Convolutional Neural Networks (FCNNs) with contracting and expansive paths (e.g. encoder and decoder) have shown prominence in various medical image segmentation applications during the recent years. In these architectures, the encoder plays an integral role by learning global contextual representations which will be further utilized for semantic output prediction by the decoder. Despite their success, the locality of convolutional layers , as the main building block of FCNNs limits the capability of learning long-range spatial dependencies in such networks. Inspired by the recent success of transformers in Natural Language Processing (NLP) in long-range sequence learning, we reformulate the task of volumetric (3D) medical image segmentation as a sequence-to-sequence prediction problem. In particular, we introduce a novel architecture, dubbed as UNEt TRansformers (UNETR), that utilizes a pure transformer as the encoder to learn sequence representations of the input volume and effectively capture the global multi-scale information. The transformer encoder is directly connected to a decoder via skip connections at different resolutions to compute the final semantic segmentation output. We have extensively validated the performance of our proposed model across different imaging modalities(i.e. MR and CT) on volumetric brain tumour and spleen segmentation tasks using the Medical Segmentation Decathlon (MSD) dataset, and our results consistently demonstrate favorable benchmarks.

Keywords: Transformers · Semantic Segmentation · Deep Learning

1 Introduction

Medical image segmentation plays an integral role in numerous clinical diagnosis methods and is often the first step for quantified analysis of anatomical structures. Since the advent of deep learning, FCNNs and in particular encoder-decoder architectures [19,13,14,12] have achieved state-of-the-art results in various medical semantic segmentation tasks [1,22,11]. In a typical U-Net [21] architecture, the encoder is responsible for learning global contextual representations by gradually downsampling the extracted features, while the decoder upsamples the extracted representations to the input resolution for pixel/voxel-wise semantic prediction. In addition, skip connections merge the output of the encoder with decoder at different resolutions, hence allowing for recovering spatial information that is lost during downsampling.

Although such FCNN-based approaches have powerful representation learning capabilities, their performance in learning long-range dependencies is limited to their localized receptive fields. As a result, such a deficiency in capturing multi-scale contextual information leads to sub-optimal segmentation of structures with variable shapes and scales (e.g. brain lesions with different sizes). Several efforts have tried to mitigate this issue by employing atrous convolutional layers [4,15,10]. However, due to the locality of CNNs, their receptive fields are still limited to a small region.

In the NLP domain, transformer-based models [24,6] have achieved state-of-the-art benchmarks in various tasks. The self-attention mechanism in the transformers enables them to dynamically highlight the important features of word sequences and learn its long-range dependencies. This notion has recently been extended to computer vision by the introduction of Visual Transformer (ViT) [7]. In ViT, an image is represented as a sequence of patch embeddings that will be used for direct prediction of class labels for image classification.

In this work, we propose to leverage the power of transformers for volumetric medical image segmentation and introduce a novel architecture dubbed as UNETR for this purpose. In particular, we reformulate the task of 3D segmentation as a 1D sequence-to-sequence prediction problem and use a pure transformer as the encoder to learn contextual information from the embedded input patches. The extracted representations from transformer encoder is merged with a decoder via skip connections at multiple resolutions for prediction of segmentation outputs.

We have extensively validated the effectiveness of our UNETR on brain tumour and spleen segmentation tasks in the MSD dataset [22] and our experiments demonstrate favorable performance in comparison to other models in our validation set. To the best of our knowledge, we are the first to propose a completely transformer-based encoder for volumetric medical image segmentation. Considering the prevalence of volumetric data in medical imaging and their extensive use in segmentation, we believe our UNETR paves the way for a new class of transformer-based segmentation models which can be utilized for various applications.

2 Related Work

CNN-based Segmentation Networks : Since introduction of the seminal U-Net [21], CNN-based networks have achieved state-of-the-art results on various 2D and 3D various medical image segmentation tasks [8,29,25,9,16,28]. Despite their success, a limitation of these networks is their poor performance in learning global context and long-range spatial dependencies, which can severely impact the segmentation performance for challenging tasks.

Visual Transformers : Visual transformers have recently gained traction for various computer vision tasks. Dosovitskiy *et al.* [7] demonstrated state-of-the-art performance on image classification datasets by large-scale pretraining and

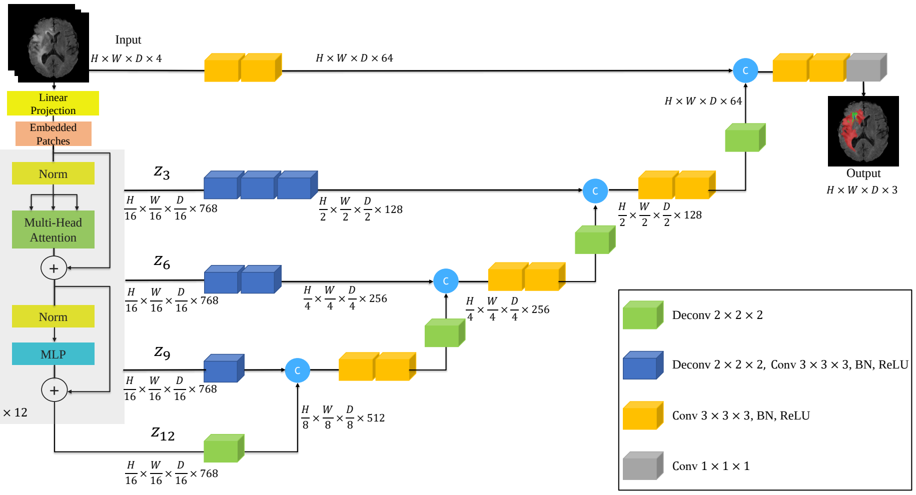
Input
$H \times W \times D \times 4$
$H \times W \times D \times 64$
Linear Projection
Embedded Patches
Norm
Multi-Head Attention
+
Norm
MLP
+
× 12
z_3
$\frac{H}{16} \times \frac{W}{16} \times \frac{D}{16} \times 768$
$\frac{H}{2} \times \frac{W}{2} \times \frac{D}{2} \times 128$
z_6
$\frac{H}{16} \times \frac{W}{16} \times \frac{D}{16} \times 768$
$\frac{H}{4} \times \frac{W}{4} \times \frac{D}{4} \times 256$
z_9
$\frac{H}{16} \times \frac{W}{16} \times \frac{D}{16} \times 768$
$\frac{H}{8} \times \frac{W}{8} \times \frac{D}{8} \times 512$
z_{12}
$\frac{H}{16} \times \frac{W}{16} \times \frac{D}{16} \times 768$
C
$\frac{H}{4} \times \frac{W}{4} \times \frac{D}{4} \times 256$
C
$\frac{H}{2} \times \frac{W}{2} \times \frac{D}{2} \times 128$
C
$H \times W \times D \times 64$
C
Output
$H \times W \times D \times 3$
Deconv 2 × 2 × 2
Deconv 2 × 2 × 2, Conv 3 × 3 × 3, BN, ReLU
Conv 3 × 3 × 3, BN, ReLU
Conv 1 × 1 × 1

Fig. 1. Overview of the UNETR architecture. We extract sequence representations of different layers in the transformer and merge them with the decoder via skip connections. Output sizes demonstrated for patch dimension $N = 16$ and embedding size $C = 768$.

fine-tuning of a pure transformer. In object detection, end-to-end transformer-based models have shown prominence on several benchmarks [2,30]. Recently, a few efforts [27,3,23,26] have explored the possibility of using transformer-based models for the task of 2D image segmentation. Chen *et al.* [3] proposed a 2D methodology for multi-organ segmentation by employing a transformer as a layer in the bottleneck of a U-Net. In addition, Zhang *et al.* [26] proposed to use CNNs and transformers in separate streams and fuse their outputs. Valanarasu *et al.* [23] proposed a transformer-based axial attention mechanism for 2D medical image segmentation. There are three key differences between our model and these efforts: (1) our UNETR is tailored for 3D segmentation and directly utilizes volumetric data; (2) our UNETR employs the transformer as the main encoder of a segmentation network and directly connects it to the decoder via skip connections, as opposed to using it as an attention layer within the segmentation network; (3) our UNETR does not rely on a backbone CNN for generating the input sequences and directly utilizes the tokenized patches.

3 Methodology

3.1 Architecture

We have presented an overview of the proposed model in Fig. 1. The UNETR utilizes a contracting-expansive pattern consisting of a stack of transformers as the encoder which is connected to a decoder via skip connections. We begin by first describing the working mechanism of our transformer encoder. As commonly used in NLP, the transformers operate on 1D sequence of input embeddings. In our UNETR, we create a 1D sequence of a 3D input volume $\mathbf{x} \in \mathbb{R}^{H\times W\times D\times C}$ by dividing it into flattened uniform non-overlapping patches $\mathbf{x}_v \in \mathbb{R}^{L\times C\times N^3}$ where (N, N, N) denotes the dimension of each patch and $L = (H \times W \times D)/N^3$ is the length of the sequence.

Subsequently, we use a linear layer to project the flattened patches into a K dimensional embedding space, which remains constant throughout the transformer. Furthermore, in order to preserve the spatial information of the extracted patches, we add a 1D learnable positional embedding $\mathbf{E}_{pos} \in \mathbb{R}^{L\times D}$ to the projected patch embedding $\mathbf{E} \in \mathbb{R}^{L^2\times C\times K}$ according to

$$\mathbf{z}_0 = [\mathbf{x}_v^1\mathbf{E}; \mathbf{x}_v^2\mathbf{E}; ...; \mathbf{x}_v^L\mathbf{E}] + \mathbf{E}_{pos}, \tag{1}$$

After the embedding layer, we utilize a stack of transformer blocks [24,7] comprising of multiheaded self-attention (MSA) and multilayer perceptron (MLP) sublayers according to

$$\mathbf{z'}_i = \mathrm{MSA}(\mathrm{Norm}(\mathbf{z}_{i-1})) + \mathbf{z}_{i-1}, \tag{2}$$

$$\mathbf{z}_i = \mathrm{MLP}(\mathrm{Norm}(\mathbf{z'}_i)) + \mathbf{z'}_i, \tag{3}$$

Where Norm represents layer normalization, MLP comprises of two linear layers with GELU activation functions and i is the intermediate block identifier ranging from 1 to $T = 12$ total blocks in our current setting. A MSA block comprises of n parallel self-attention (SA) heads. The (SA) block is a parameterized function that learns the similarity between two elements in the input sequence ($\mathbf{z}$) and their set of query ($\mathbf{q}$) and key ($\mathbf{k}$) representations. Thus, the output of (SA) is computed as follows

$$\mathrm{SA}(\mathbf{z}) = \mathrm{Softmax}(\frac{qk^\top}{\sqrt{C_h}})v, \tag{4}$$

Where $\mathbf{v}$ denotes the values in the input sequence and $C_h = C/n$ is a scaling factor. Furthermore, the output of MSA is defined as

$$\mathrm{MSA}(\mathbf{z}) = [\mathrm{SA}_1(z); \mathrm{SA}_2(z); ...; \mathrm{SA}_n(z)]\mathbf{W}_{msa}. \tag{5}$$

Where $\mathbf{W}_{msa}$ represents the learnable weight matrices of different heads (SA).

Inspired by UNet-like architectures, where features from multiple resolutions of the encoder are merged with the decoder, we extract sequence representation $\mathbf{z}_i$ ($i \in \{3, 6, 9, 12\}$), with size $\frac{H\times W\times D}{N^3} \times C$, from the transformer and reshape them into a $\frac{H}{N} \times \frac{W}{N} \times \frac{D}{N} \times C$ tensor. A representation in our definition is in the embedding space if it has been reshaped as an output of the transformer and has a feature size of C (i.e. transformer's embedding size). Consequently, we project the reshaped tensor from the embedding space into the input space by utilizing consecutive $3 \times 3 \times 3$ convolutional layers that are followed by batch normalization (See Fig. 1 for details).

At the bottleneck of our encoder (i.e. output of transformer's last layer), we apply a deconvolutional layer to the transformed feature map to increase its resolution by a factor of 2. We then concatenate the resized feature map with the feature map of the previous transformer output (e.g. $\mathbf{z}_9$), feed them into consecutive $3 \times 3 \times 3$ convolutional layers and upsample the output using a deconvolutional layer. This process is repeated for all the other subsequent layers up to the original input resolution where the final output is fed into a $1 \times 1 \times 1$ convolutional layer with a softmax activation function to generate pixel-wise semantic predictions.

3.2 Loss Function

Our loss function is a combination of dice [18] and cross entropy terms that can be computed in a voxel-wise manner according to

$$\mathcal{L} = 1 - \frac{2}{J} \sum_{j=1}^{J} \frac{\sum_{i=1}^{I} G_{i,j} Y_{i,j}}{\sum_{i=1}^{I} G_{i,j}^2 + \sum_{i=1}^{I} Y_{i,j}^2} - \frac{1}{I} \sum_{i=1}^{I} \sum_{j=1}^{J} G_{i,j} \log Y_{i,j}. \qquad (6)$$

where I is the number of voxels; J is the number of classes; $Y_{i,j}$ and $G_{i,j}$ denote the probability output and one-hot encoded ground truth for class j at voxel i, respectively.

4 Experiments

4.1 Datasets

To cover various objects and image modalities, the datasets of task 1 (brain tumour MRI segmentation) and task 9 (spleen CT segmentation) from MSD challenge [22] are adopted for experiments with our own data split of 5-fold cross validation. For task 1, the entire training set of 484 multi-modal multi-site MRI data (FLAIR, T1w, T1gd, T2w) with ground truth labels of gliomas segmentation necrotic/active tumour and oedema is utilized for model training. The resolution/spacing of task 1 is uniformly $1.0 \times 1.0 \times 1.0$ mm^3. For task 9, 41 CT volumes with spleen body annotation are used. The resolution/spacing of volumes in task 9 ranges from $0.613 \times 0.613 \times 1.50$ mm^3 to $0.977 \times 0.977 \times 8.0$ mm^3. All volumes are re-sampled into the isotropic voxel spacing 1.0 mm during pre-processing.

For task 1 with MRI images, the voxel intensities are pre-processed with z-score normalization. For task 9 with CT images, the voxel intensities of the images are normalized to the range $[0, 1]$ according to 5th and 95th percentile of overall foreground intensities. Furthermore, the problem of task 1 is formulated as a 3-class segmentation task with 4-channel input whereas task 9 is formulated as a binary segmentation task (foreground and background) with single-channel input. We randomly sample the input images with volume sizes of $[128, 128, 128]$ and $[96, 96, 96]$ for tasks 1 and 9 respectively. The random patches of foreground/background are sampled at ratio $1 : 1$.

Fold	Split-1	Split-2	Split-3	Split-4	Split-5	DSC1	DSC2	DSC3	Avg.
VNet [18]	64.83	67.28	65.23	65.2	66.34	75.96	54.99	66.38	65.77
AHNet [17]	65.78	69.31	65.16	65.05	67.84	75.8	57.58	66.50	66.63
Att-UNet [20]	66.39	70.18	65.39	66.11	67.29	75.29	57.11	68.81	67.07
UNet [5]	67.20	69.11	66.84	66.95	68.16	75.03	57.87	70.06	67.65
SegResNet [19]	69.62	71.84	67.86	68.52	70.43	76.37	59.56	73.03	69.65
UNETR	**70.79**	**73.70**	**70.12**	**72.10**	**72.38**	**79.00**	**60.62**	**75.82**	**71.81**

Table 1. Cross validation results of brain tumour Segmentation task. For each split, we provide the average dice score of three classes. DSC1, DSC2 and DSC3 denote average dice scores for the Whole Tumour (WT), Enhancing Tumour (ET) and Tumour Core (TC) across all folds respectively.

Fold	Split-1	Split-2	Split-3	Split-4	Split-5	Avg.
VNet [18]	94.78	92.08	95.54	94.73	95.03	94.43
AHNet [17]	94.23	92.10	94.56	94.39	94.11	93.87
Att-UNet [20]	93.16	92.59	95.08	94.75	95.81	94.27
UNet [5]	92.83	92.83	95.76	95.01	96.27	94.54
SegResNet [19]	95.66	92.00	95.79	94.19	95.53	94.63
UNETR	**95.95**	**94.01**	**96.37**	**95.89**	**96.91**	**95.82**

Table 2. Cross validation results of spleen segmentation task. For each split, we provide the average dice score of fore-ground class.

4.2 Implementation Details

The UNETR is implemented in Pytorch[1] and MONAI[2]. The model was trained on a NVIDIA V100 32GB GPU and an Intel® Core™ i7-7800X CPU @ 3.50GHz × 12. All models were trained with a batch size of 2 and using the Adam optimization algorithm with initial learning rate of 0.0001 for $25,000$ iterations. Using a fixed split for all experiments, we have used five fold cross-validation and evaluated the performance of our model by using Dice-Sørensen score (DSC). We have used a dimension of $16 \times 16 \times 16$ for generating the input patches, and $T = 12$ transformer blocks with embedding size of $C = 768$ as the encoder of UNETR. We did not use any pretrained transformer model(e.g. ViT on ImageNet) since pretraining did not show any performance improvement.

4.3 Quantitative Evaluations

In Table 1, we compare the performance of the UNETR against baseline CNN-based networks for the task of brain tumour segmentation. Our UNETR outperforms the closest baseline by 2.28% on average over all semantic classes. In particular, the UNETR performs considerably better in segmenting Tumour Core (TC). In Table 2, We compare the performance of UNETR against baselines for

[1] http://pytorch.org/

[2] https://monai.io/

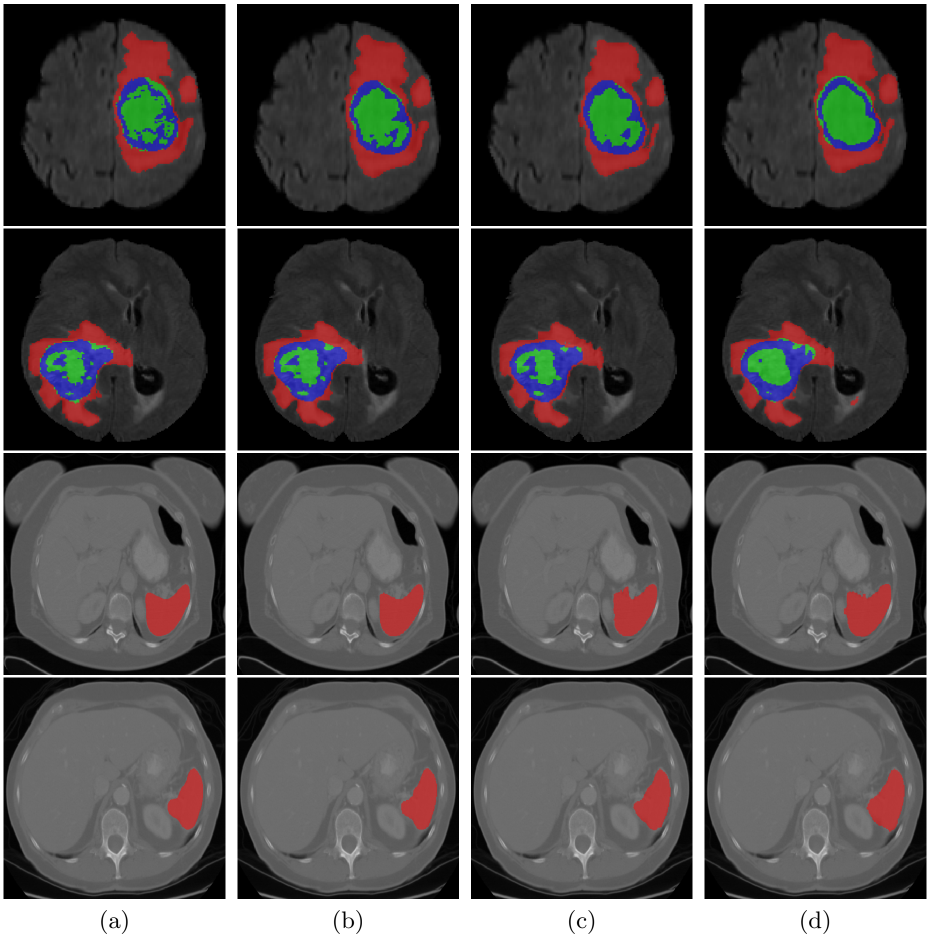

Fig. 2. (a) Ground Truth. Outputs of : (b) UNETR. (c) SegResNet. (d) UNet.

the task of spleen segmentation. Similarly, the UNETR outperforms the closest baselines by least 1.11%. Furthermore, in order to allow for a fair comparison, we did not compare against external models on MSD test set, since leveraging ensembles, commonly used for boosting the test time performance and different training conditions can significantly alter the benchmarks.

4.4 Qualitative Results

In Fig. 2, we present qualitative results of our model's segmentation outputs as well as other baselines. For brain tumour segmentation, our model demonstrates better performance in capturing fine-grained structural details of tumours.

5 Ablation

Decoder Choice In Table 3, we evaluate the effectiveness of the proposed architecture by using our UNETR encoder with different decoders. In these experiments, we employed the encoder of UNETR (i.e. stack of transformers) but replaced the decoder with 3D counterpart of naive and progressive upsampling as well as multi-scale aggregation [27]. We observe that although MLA marginally performs better than other decoders, they still yield sub-optimal results. Our UNETR with its proposed decoder outperforms MLA by 6.09% and 2.66% on brain tumor and spleen segmentation tasks.

Decoder	NUP	PUP	MLA	UNETR
Brain(DSC,Average)	64.43	65.18	65.72	**71.81**
Spleen(DSC)	92.85	93.07	93.16	**95.82**

Table 3. Effect of decoder on segmentation performance. NUP, PUP and MLA denote Naive UpSampling, Progressive UpSampling and Multi-scale Aggregation.

Patch Dimension A lower input patch dimension leads to a higher sequence length since it is inversely correlated to the cube of the dimension. As shown in Table 4, our experiments demonstrate that decreasing the dimensions leads to consistently improved performance in both tasks. We did not experiment with lower dimensions due to memory constrains.

Dimension(N)	Spleen(DSC)	Brain(DSC,Avg)	Brain(DSC1)	Brain(DSC2)	Brain(DSC3)
16	**95.82**	**71.81**	**79.06**	**60.62**	**76.01**
32	95.12	70.98	77.83	60.01	75.12

Table 4. Effect of patch dimension on segmentation performance.

6 Conclusion

In this paper, we introduced a novel transformer-based architecture for volumetric semantic segmentation of medical images and proposed to reformulate it as a 1D sequence-to-sequence prediction. We proposed to use a pure transformer encoder to increase the model's capability for learning long-range dependencies and effectively capturing global contextual representation at multiple scales.

We validated the effectiveness of our UNETR on volumetric brain tumour and spleen segmentation tasks of MSD dataset in CT and MR image modalities and our benchmarks consistently demonstrated favorable performance on these tasks. Our proposed UNETR lays the foundation for a new class of transformer-based models for medical image segmentation. Although the UNETR is primarily designed for 3D applications, an extension for 2D applications is straightforward and can be explored in future efforts.

References

1. Bakas, S., Reyes, M., et Int, Menze, B.: Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. In: arXiv:1811.02629 (2018)
2. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. arXiv preprint arXiv:2005.12872 (2020)
3. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
4. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence **40**(4), 834–848 (2017)
5. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: International conference on medical image computing and computer-assisted intervention. pp. 424–432. Springer (2016)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
8. Dou, Q., Chen, H., Jin, Y., Yu, L., Qin, J., Heng, P.A.: 3d deeply supervised network for automatic liver segmentation from ct volumes. In: International conference on medical image computing and computer-assisted intervention. pp. 149–157. Springer (2016)
9. Gibson, E., Giganti, F., Hu, Y., Bonmati, E., Bandula, S., Gurusamy, K., Davidson, B., Pereira, S.P., Clarkson, M.J., Barratt, D.C.: Automatic multi-organ segmentation on abdominal ct with dense v-networks. IEEE transactions on medical imaging **37**(8), 1822–1834 (2018)
10. Gu, Z., Cheng, J., Fu, H., Zhou, K., Hao, H., Zhao, Y., Zhang, T., Gao, S., Liu, J.: Ce-net: Context encoder network for 2d medical image segmentation. IEEE transactions on medical imaging **38**(10), 2281–2292 (2019)
11. Heller, N., Sathianathen, N., Kalapara, A., Walczak, E., Moore, K., Kaluzniak, H., Rosenberg, J., Blake, P., Rengel, Z., Oestreich, M., et al.: The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes. arXiv preprint arXiv:1904.00445 (2019)
12. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2021)

13. Isensee, F., Maier-Hein, K.H.: An attempt at beating the 3d u-net. arXiv preprint arXiv:1908.02182 (2019)
14. Jin, Q., Meng, Z., Sun, C., Cui, H., Su, R.: Ra-unet: A hybrid deep attention-aware network to extract liver and tumor in ct scans. Frontiers in Bioengineering and Biotechnology **8**, 1471 (2020)
15. Li, W., Wang, G., Fidon, L., Ourselin, S., Cardoso, M.J., Vercauteren, T.: On the compactness, efficiency, and representation of 3d convolutional networks: brain parcellation as a pretext task. In: International conference on information processing in medical imaging. pp. 348–360. Springer (2017)
16. Li, X., Chen, H., Qi, X., Dou, Q., Fu, C.W., Heng, P.A.: H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. IEEE transactions on medical imaging **37**(12), 2663–2674 (2018)
17. Liu, S., Xu, D., Zhou, S.K., Pauly, O., Grbic, S., Mertelmeier, T., Wicklein, J., Jerebko, A., Cai, W., Comaniciu, D.: 3d anisotropic hybrid network: Transferring convolutional features from 2d images to 3d anisotropic volumes. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 851–858. Springer (2018)
18. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV). pp. 565–571. IEEE (2016)
19. Myronenko, A.: 3D MRI brain tumor segmentation using autoencoder regularization. In: BrainLes, Medical Image Computing and Computer Assisted Intervention (MICCAI). pp. 311–320. LNCS, Springer (2018)
20. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al.: Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)
21. Ronneberger, O., P.Fischer, Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Proc. MICCAI. LNCS, vol. 9351, pp. 234–241 (2015)
22. Simpson, A.L., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., Van Ginneken, B., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., et al.: A large annotated medical image dataset for the development and evaluation of segmentation algorithms. arXiv preprint arXiv:1902.09063 (2019)
23. Valanarasu, J.M.J., Oza, P., Hacihaliloglu, I., Patel, V.M.: Medical transformer: Gated axial-attention for medical image segmentation. arXiv preprint arXiv:2102.10662 (2021)
24. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
25. Yu, L., Yang, X., Chen, H., Qin, J., Heng, P.A.: Volumetric convnets with mixed residual connections for automated prostate segmentation from 3d mr images. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 31 (2017)
26. Zhang, Y., Liu, H., Hu, Q.: Transfuse: Fusing transformers and cnns for medical image segmentation. arXiv preprint arXiv:2102.08005 (2021)
27. Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. arXiv preprint arXiv:2012.15840 (2020)
28. Zhou, Z., Siddiquee, M.M.R., 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. Springer (2018)

29. Zhu, Q., Du, B., Turkbey, B., Choyke, P.L., Yan, P.: Deeply-supervised cnn for prostate segmentation. In: 2017 international joint conference on neural networks (IJCNN). pp. 178–184. IEEE (2017)
30. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)