UNETR: Transformers for 3D Medical Image Segmentation

Ali Hatamizadeh
NVIDIA

Yucheng Tang
Vanderbilt University

Vishwesh Nath
NVIDIA

Dong Yang
NVIDIA

Andriy Myronenko
NVIDIA

Bennett Landman
Vanderbilt University

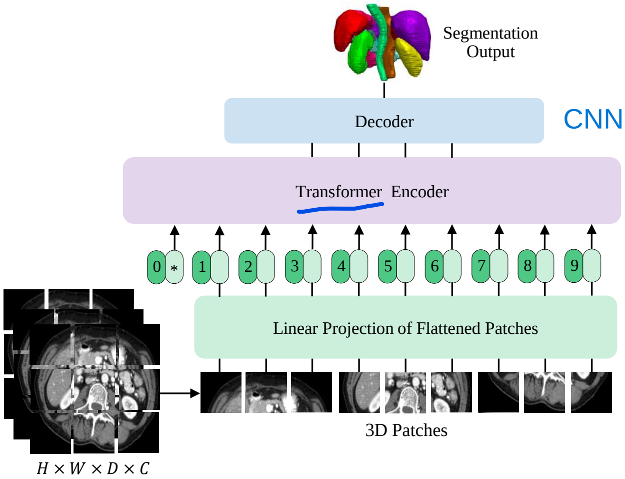
Holger R. Roth
NVIDIA

Daguang Xu
NVIDIA

Unet uses conv blocks as the "encoder" part, here they propose using a transformer as the "encoder" to create the embeddings where the decoder is still Conv blocks

Abstract

Fully Convolutional Neural Networks (FCNNs) with contracting and expanding paths have shown prominence for the majority of medical image segmentation applications since the past decade. In FCNNs, the encoder plays an integral role by learning both global and local features and contextual representations which can be utilized for semantic output prediction by the decoder. Despite their success, the locality of convolutional layers in FCNNs, limits the capability of learning long-range spatial dependencies. Inspired by the recent success of transformers for 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. We introduce a novel architecture, dubbed as UNEt TRansformers (UNETR), that utilizes a transformer as the encoder to learn sequence representations of the input volume and effectively capture the global multi-scale information, while also following the successful "U-shaped" network design for the encoder and decoder. The transformer encoder is directly connected to a decoder via skip connections at different resolutions to compute the final semantic segmentation output. We have validated the performance of our method on the Multi Atlas Labeling Beyond The Cranial Vault (BTCV) dataset for multi-organ segmentation and the Medical Segmentation Decathlon (MSD) dataset for brain tumor and spleen segmentation tasks. Our benchmarks demonstrate new state-of-the-art performance on the BTCV leaderboard.
Code: *https://monai.io/research/unetr*


Segmentation Output
Decoder
Transformer Encoder
0
*
1
2
3
4
5
6
7
8
9
Linear Projection of Flattened Patches
3D Patches
$H \times W \times D \times C$

CNN decoder

Figure 1. Overview of UNETR. Our proposed model consists of a transformer encoder that directly utilizes 3D patches and is connected to a CNN-based decoder via skip connection.

1. Introduction

Image segmentation plays an integral role in quantitative medical image analysis as it is often the first step for analysis of anatomical structures [33]. Since the advent of deep learning, FCNNs and in particular "U-shaped" encoder-decoder architectures [22, 23, 21] have achieved state-of-the-art results in various medical semantic segmentation tasks [2, 38, 19]. In a typical U-Net [36] 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.

skip connections refers to the concatenation (not residual)

Although such FCNN-based approaches have powerful representation learning capabilities, their performance in learning long-range dependencies is limited to their localized receptive fields [20, 35]. As a result, such a deficiency in capturing multi-scale information leads to sub-optimal segmentation of structures with variable shapes and scales (e.g. brain lesions with different sizes). Several efforts have used atrous convolutional layers [9, 27, 18] to enlarge the receptive fields. However, locality of the receptive fields in convolutional layers still limits their learning capabilities to relatively small regions. Combining self-attention modules with convolutional layers [45, 50, 16] has been proposed to improve the non-local modeling capability.

In Natural Language Processing (NLP), transformer-based models [42, 13] achieve state-of-the-art benchmarks in various tasks. The self-attention mechanism of transformers allows to dynamically highlight the important features of word sequences. Additionally, in computer vision, using transformers as a backbone encoder is beneficial due to their great capability of modeling long-range dependencies and capturing global context [14, 4]. Specifically, unlike the local formulation of convolutions, transformers encode images as a sequence of 1D patch embeddings and utilize self-attention modules to learn a weighted sum of values that are calculated from hidden layers. As a result, this flexible formulation allows to effectively learn the long-range information. Furthermore, Vision Transformer (ViT) [14] and its variants have shown excellent capabilities in learning pre-text tasks that can be transferred to down-stream applications [40, 6, 3].

transformer encoder turns image patches into a vector embedding representaiton

BUT also, the input to the transformer encoder are 1D vectors being the flattened image patch after going through linear layer

In this work, we propose to leverage the power of transformers for volumetric medical image segmentation and introduce a novel architecture dubbed as UNEt TRansformers (UNETR). In particular, we reformulate the task of 3D segmentation as a 1D sequence-to-sequence prediction problem and use a transformer as the encoder to learn contextual information from the embedded input patches. The extracted representations from the transformer encoder are merged with the CNN-based decoder via skip connections at multiple resolutions to predict the segmentation outputs. Instead of using transformers in the decoder, our proposed framework uses a CNN-based decoder. This is due to the fact that transformers are unable to properly capture localized information, despite their great capability of learning global information.

We validate the effectiveness of our method on 3D CT and MRI segmentation tasks using Beyond the Cranial Vault (BTCV) [26] and Medical Segmentation Decathlon (MSD) [38] datasets. In BTCV dataset, UNETR achieves new state-of-the-art performance on both Standard and Free Competition sections on its leaderboard. UNETR outperforms the state-of-the-art methodologies on both brain tumor and spleen segmentation tasks in MSD dataset.

our main contributions of this work are as follows::

- We propose a novel transformer-based model for volumetric medical image segmentation.
- To this end, we propose a novel architecture in which (1) a transformer encoder directly utilizes the embedded 3D volumes to effectively capture long-range dependencies; (2) a skip-connected decoder combines the extracted representations at different resolutions and predicts the segmentation output.

it seems their encoder process 3D image patches (does not process 2Dslices like i did naively)

- We validate the effectiveness of our proposed model for different volumetric segmentation tasks on two public datasets: BTCV [26] and MSD [38]. UNETR achieves new state-of-the-art performance on *leaderboard* of BTCV dataset and outperforms competing approaches on the MSD dataset.

2. Related Work

CNN-based Segmentation Networks : Since the introduction of the seminal U-Net [36], CNN-based networks have achieved state-of-the-art results on various 2D and 3D various medical image segmentation tasks [15, 54, 49, 17, 28]. For volume-wise segmentation, tri-planar architectures are sometimes used to combine three-view slices for each voxel, also known for 2.5D methods [28, 29, 46]. In contrast, 3D approaches directly utilize the full volumetric image represented by a sequence of 2D slices or modalities. The intuition of employing varying sizes was followed by multi-scan, multi-path models [24, 25, 8] to capture downsampled features of the image. In addition, to exploit 3D context and to cope with limitation of computational resource, researchers investigated hierarchical frameworks.

Some efforts proposed to extract features at multiple scales or assembled frameworks [21]. Roth *et al.* [37] proposed a multi-scale framework to obtain varying resolution information in pancreas segmentation. These methods provide pioneer studies of 3D medical image segmentation at multiple levels, which reduces problems in spatial context and low-resolution condition. 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.

Vision Transformers : Vision transformers have recently gained traction for computer vision tasks. Dosovitskiy *et al.* [14] demonstrated state-of-the-art performance on image classification datasets by large-scale pre-training and fine-tuning of a pure transformer. In object detection, end-to-end transformer-based models have shown prominence on several benchmarks [5, 55]. Recently, hierarchical vision transformers with varying resolutions and spatial embeddings [30, 44, 12, 48] have been proposed. These methodologies gradually decrease the resolution of features in the transformer layers and utilize sub-sampled attention modules. Unlike these approaches, the size of representation in UNETR encoder remains fixed in all transformer layers. However, as described in Sec. 3, deconvolutional and convolutional operations are used to change the resolution of extracted features.

Recently, multiple methods were proposed that explore the possibility of using transformer-based models for the task of 2D image segmentation [52, 7, 41, 51]. Zheng *et al.* [52] introduced the SETR model in which a pre-trained transformer encoder with different variations of CNN-based decoders were proposed for the task of semantic segmentation. Chen *et al.* [7] proposed a methodology for multi-organ segmentation by employing a transformer as an additional layer in the bottleneck of a U-Net architecture. Zhang *et al.* [51] proposed to use CNNs and transformers in separate streams and fuse their outputs. Valanarasu *et al.* [41] proposed a transformer-based axial attention mechanism for 2D medical image segmentation. There are key differences between our model and these efforts: (1) UNETR is tailored for 3D segmentation and directly utilizes volumetric data; (2) 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) UNETR does not rely on a backbone CNN for generating the input sequences and directly utilizes the tokenized patches.

For 3D medical image segmentation, Xie *et al.* [47] proposed a framework that utilizes a backbone CNN for feature extraction, a transformer to process the encoded representation and a CNN decoder for predicting the segmentation outputs. Similarly, Wang *et al.* [43] proposed to use a transformer in the bottleneck of a 3D encoder-decoder CNN for the task of semantic brain tumor segmentation. In contrast to these approaches, our method directly connects the encoded representation from the transformer to decoder by using skip connections.

3. Methodology

3.1. Architecture

We have presented an overview of the proposed model in Fig. 2. UNETR utilizes a contracting-expanding pattern consisting of a stack of transformers as the encoder which is connected to a decoder via skip connections. As commonly used in NLP, the transformers operate on 1D sequence of input embeddings. Similarly, we create a 1D sequence of a 3D input volume $\mathbf{x} \in \mathbb{R}^{H \times W \times D \times C}$ with resolution (H,W,D) and C input channels by dividing it into flattened uniform non-overlapping patches $\mathbf{x}_v \in \mathbb{R}^{N \times (P^3 . C)}$ where (P,P,P) denotes the resolution of each patch and $N = (H \times W \times D)/P^3$ is the length of the sequence.

Subsequently, we use a linear layer to project the patches into a K dimensional embedding space, which remains constant throughout the transformer layers. In order to preserve the spatial information of the extracted patches, we add a 1D learnable positional embedding $\mathbf{E}_{pos} \in \mathbb{R}^{N \times K}$ to the projected patch embedding $\mathbf{E} \in \mathbb{R}^{(P^3 . C) \times K}$ according to

$$\mathbf{z}_0 = [\mathbf{x}_v^1\mathbf{E}; \mathbf{x}_v^2\mathbf{E}; ...; \mathbf{x}_v^N\mathbf{E}] + \mathbf{E}_{pos}, \qquad (1)$$

Note that the learnable `[class]` token is not added to the sequence of embeddings since our transformer backbone is designed for semantic segmentation. After the embedding layer, we utilize a stack of transformer blocks [42, 14] comprising of multi-head self-attention (MSA) and multilayer perceptron (MLP) sublayers according to

$$\mathbf{z}'_i = \text{MSA}(\text{Norm}(\mathbf{z}_{i-1})) + \mathbf{z}_{i-1}, \quad i = 1...L, \qquad (2)$$

$$\mathbf{z}_i = \text{MLP}(\text{Norm}(\mathbf{z}'_i)) + \mathbf{z}'_i, \quad i = 1...L, \qquad (3)$$

where Norm() denotes layer normalization [1], MLP comprises of two linear layers with GELU activation functions, i is the intermediate block identifier, and L is the number of transformer layers.

A MSA sublayer comprises of n parallel self-attention (SA) heads. Specifically, the SA block, is a parameterized function that learns the mapping between a query ($\mathbf{q}$) and the corresponding key ($\mathbf{k}$) and value ($\mathbf{v}$) representations in a sequence $\mathbf{z} \in \mathbb{R}^{N \times K}$. The attention weights (A) are computed by measuring the similarity between two elements in $\mathbf{z}$ and their key-value pairs according to

$$A = \text{Softmax}(\frac{\mathbf{q}\mathbf{k}^\top}{\sqrt{K_h}}), \qquad (4)$$

where $K_h = K/n$ is a scaling factor for maintaining the number of parameters to a constant value with different values of the key $\mathbf{k}$. Using the computed attention weights, the output of SA for values $\mathbf{v}$ in the sequence $\mathbf{z}$ is computed as

$$\text{SA}(\mathbf{z}) = A\mathbf{v}, \qquad (5)$$

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

$$\text{MSA}(\mathbf{z}) = [\text{SA}_1(\mathbf{z}); \text{SA}_2(\mathbf{z}); ...; \text{SA}_n(\mathbf{z})]\mathbf{W}_{msa}, \qquad (6)$$

where $\mathbf{W}_{msa} \in \mathbb{R}^{n.K_h \times K}$ represents the multi-headed trainable parameter weights.

Inspired by architectures that are similar to U-Net [36], where features from multiple resolutions of the encoder are merged with the decoder, we extract a sequence representation $\mathbf{z}_i$ ($i \in \{3,6,9,12\}$), with size $\frac{H \times W \times D}{P^3} \times K$, from the transformer and reshape them into a $\frac{H}{P} \times \frac{W}{P} \times \frac{D}{P} \times K$ tensor. A representation in our definition is in the embedding space after it has been reshaped as an output of the transformer with feature size of K (i.e. transformer's embedding size). Furthermore, as shown in Fig. 2, at each resolution we project the reshaped tensors from the embedding space into the input space by utilizing consecutive $3 \times 3 \times 3$ convolutional layers that are followed by normalization layers.

just realized but pytorch does have a Conv3D method (so far only have used Conv2D and did not understand Conv3D)

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

as done in unet

this was one of my confusions after using a pretrained encoder. Here it seems since they're training the transformer themselves, they did not include a cls token (for the model to learn this you would have to add this to your vocab and add it as the first token in the sequence, but here they indicated they did not add it

SO, when using a pretrained transforer encoder from say hugging face which by default i think has a learned cls token, you have to omit this in the decoder i guess

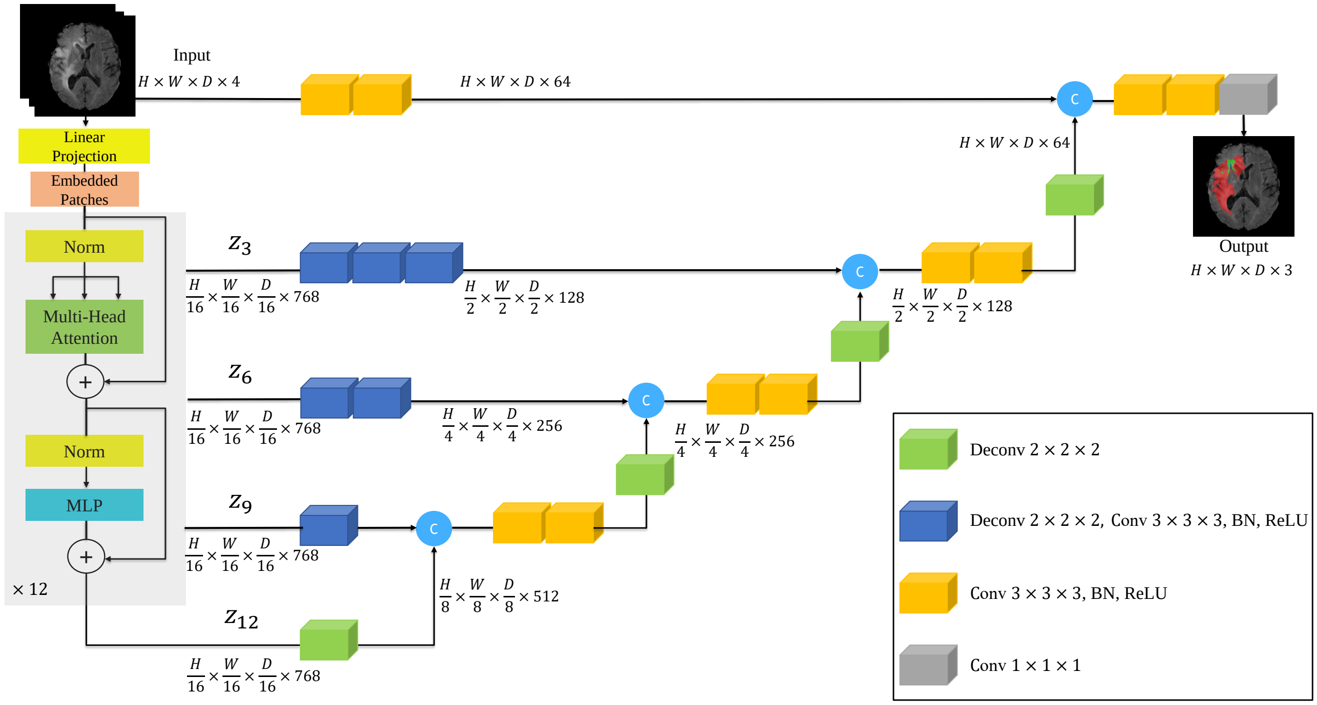

Input
$H \times W \times D \times 4$
$H \times W \times D \times 64$
Linear Projection
Embedded Patches
Norm
Multi-Head Attention
+
Norm
MLP
+
× 12
z_3
$\frac{H}{16} \times \frac{W}{16} \times \frac{D}{16} \times 768$
$\frac{H}{2} \times \frac{W}{2} \times \frac{D}{2} \times 128$
C
$\frac{H}{2} \times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

Figure 2. Overview of UNETR architecture. A 3D input volume (e.g. $C=4$ channels for MRI images), is divided into a sequence of uniform non-overlapping patches and projected into an embedding space using a linear layer. The sequence is added with a position embedding and used as an input to a transformer model. The encoded representations of different layers in the transformer are extracted and merged with a decoder via skip connections to predict the final segmentation. Output sizes are given for patch resolution $P=16$ and embedding size $K=768$.

feature map of the previous transformer output (e.g. $\mathbf{z}_9$), and 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 voxel-wise semantic predictions.

3.2. Loss Function

Our loss function is a combination of soft dice loss [32] and cross-entropy loss, and it can be computed in a voxel-wise manner according to

$$\begin{aligned}\mathcal{L}(G,Y) = 1 - \frac{2}{J}\sum_{j=1}^{J}\frac{\sum_{i=1}^{I}G_{i,j}Y_{i,j}}{\sum_{i=1}^{I}G_{i,j}^{2}+\sum_{i=1}^{I}Y_{i,j}^{2}} - \\ -\frac{1}{I}\sum_{i=1}^{I}\sum_{j=1}^{J}G_{i,j}\log Y_{i,j}.\end{aligned} \tag{7}$$

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 validate the effectiveness of our method, we utilize BTCV [26] and MSD [38] datasets for three different segmentation tasks in CT and MRI imaging modalities.

BTCV (CT): The BTCV dataset [26] consists of 30 subjects with abdominal CT scans where 13 organs were annotated by interpreters under supervision of clinical radiologists at Vanderbilt University Medical Center. Each CT scan was acquired with contrast enhancement in portal venous phase and consists of 80 to 225 slices with 512×512 pixels and slice thickness ranging from 1 to 6 mm. Each volume has been pre-processed independently by normalizing the intensities in the range of [-1000,1000] HU to [0,1]. All images are resampled into the isotropic voxel spacing of 1.0 mm during pre-processing. The multi-organ segmentation problem is formulated as a 13 class segmentation task with 1-channel input.

MSD (MRI/CT): For the brain tumor segmentation task, 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 tumor and oedema is utilized for model training. The voxel spacing of MRI images in this tasks is $1.0 \times 1.0 \times 1.0\ mm^3$. The voxel intensities are pre-processed with z-score normalization. The

12 organs, each result is the indidual metric performance for the organ. AVG is the mean on all of the organs

Methods	Spl	RKid	LKid	Gall	Eso	Liv	Sto	Aor	IVC	Veins	Pan	AG	Avg.
SETR NUP [52]	0.931	0.890	0.897	0.652	0.760	0.952	0.809	0.867	0.745	0.717	0.719	0.620	0.796
SETR PUP [52]	0.929	0.893	0.892	0.649	0.764	0.954	0.822	0.869	0.742	0.715	0.714	0.618	0.797
SETR MLA [52]	0.930	0.889	0.894	0.650	0.762	0.953	0.819	0.872	0.739	0.720	0.716	0.614	0.796
nnUNet [21]	0.942	0.894	0.910	0.704	0.723	0.948	0.824	0.877	0.782	0.720	0.680	0.616	0.802
ASPP [10]	0.935	0.892	0.914	0.689	0.760	0.953	0.812	0.918	0.807	0.695	0.720	0.629	0.811
TransUNet [7]	0.952	**0.927**	0.929	0.662	0.757	0.969	0.889	0.920	0.833	0.791	0.775	0.637	0.838
CoTr w/o CNN encoder [47]	0.941	0.894	0.909	0.705	0.723	0.948	0.815	0.876	0.784	0.723	0.671	0.623	0.801
CoTr* [47]	0.943	0.924	0.929	0.687	0.762	0.962	0.894	0.914	0.838	**0.796**	**0.783**	0.647	0.841
CoTr [47]	0.958	0.921	0.936	0.700	0.764	0.963	0.854	**0.920**	0.838	0.787	0.775	0.694	0.844
UNETR	**0.968**	0.924	**0.941**	**0.750**	**0.766**	**0.971**	**0.913**	0.890	**0.847**	0.788	0.767	**0.741**	**0.856**
RandomPatch [39]	0.963	0.912	0.921	0.749	0.760	0.962	0.870	0.889	0.846	0.786	0.762	0.712	0.844
PaNN [53]	0.966	0.927	0.952	0.732	0.791	0.973	0.891	0.914	0.850	0.805	0.802	0.652	0.854
nnUNet-v2 [21]	0.972	0.924	**0.958**	0.780	0.841	0.976	0.922	0.921	0.872	0.831	0.842	0.775	0.884
nnUNet-dys3 [21]	0.967	0.924	0.957	0.814	0.832	0.975	0.925	0.928	0.870	0.832	**0.849**	0.784	0.888
UNETR	**0.972**	**0.942**	0.954	**0.825**	**0.864**	**0.983**	**0.945**	**0.948**	**0.890**	**0.858**	0.799	**0.812**	**0.891**

i guess these are dice scores

Table 1. Quantitative comparisons of segmentation performance in BTCV test set. Top and bottom sections represent the benchmarks of Standard and Free Competitions respectively. Our method is compared against current state-of-the-art models. All SETR [52] baselines use ViT-B-16 [14] backbone. Note: Spl: spleen, RKid: right kidney, LKid: left kidney, Gall: gallbladder, Eso: esophagus, Liv: liver, Sto: stomach, Aor: aorta IVC: inferior vena cava, Veins: portal and splenic veins, Pan: pancreas, AG: adrenal gland. All results obtained from BTCV leaderboard.

Task/Modality	Spleen Segmentation (CT)		Brain tumor Segmentation (MRI)							
Anatomy	Spleen		WT		ET		TC		All	
Metrics	Dice	HD95	Dice	HD95	Dice	HD95	Dice	HD95	Dice	HD95
UNet [36]	0.953	4.087	0.766	9.205	0.561	11.122	0.665	10.243	0.664	10.190
AttUNet [34]	0.951	4.091	0.767	9.004	0.543	10.447	0.683	10.463	0.665	9.971
SETR NUP [52]	0.947	4.124	0.697	14.419	0.544	11.723	0.669	15.192	0.637	13.778
SETR PUP [52]	0.949	4.107	0.696	15.245	0.549	11.759	0.670	15.023	0.638	14.009
SETR MLA [52]	0.950	4.091	0.698	15.503	0.554	10.237	0.665	14.716	0.639	13.485
TransUNet [7]	0.950	4.031	0.706	14.027	0.542	10.421	0.684	14.501	0.644	12.983
TransBTS [43]	-	-	0.779	10.030	0.574	9.969	0.735	8.950	0.696	9.650
CoTr w/o CNN encoder [47]	0.946	4.748	0.712	11.492	0.523	9.592	0.698	12.581	0.6444	11.221
CoTr [47]	0.954	3.860	0.746	9.198	0.557	9.447	0.748	10.445	0.683	9.697
UNETR	**0.964**	**1.333**	**0.789**	**8.266**	**0.585**	**9.354**	**0.761**	**8.845**	**0.711**	**8.822**

Table 2. Quantitative comparisons of the segmentation performance in brain tumor and spleen segmentation tasks of the MSD dataset. WT, ET and TC denote Whole Tumor, Enhancing tumor and Tumor Core sub-regions respectively.

problem of brain tumor segmentation is formulated as a 3 class segmentation task with 4-channel input.

For the spleen segmentation task, 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 of $1.0\ mm$ during pre-processing. The voxel intensities of the images are normalized to the range $[0,1]$ according to 5th and 95th percentile of overall foreground intensities. Spleen segmentation is formulated as a binary segmentation task with 1-channel input. For multi-organ and spleen segmentation tasks, we randomly sample the input images with volume sizes of $[96,96,96]$. For brain segmentation task, we randomly sample the input images with volume sizes of $[128,128,128]$. For all experiments, the random patches of foreground/background are sampled at ratio $1:1$.

4.2. Evaluation Metrics

We use Dice score and 95% Hausdorff Distance (HD) to evaluate the accuracy of segmentation in our experiments. For a given semantic class, let G_i and P_i denote the ground truth and prediction values for voxel i and G' and P' denote ground truth and prediction surface point sets respectively. The Dice score and HD metrics are defined as

$$\text{Dice}(G,P) = \frac{2\sum_{i=1}^{I} G_i P_i}{\sum_{i=1}^{I} G_i + \sum_{i=1}^{I} P_i}, \quad (8)$$

$$\begin{aligned} \text{HD}(G',P') = \max\{ & \max_{g' \in G'} \min_{p' \in P'} \|g' - p'\|, \\ & \max_{p' \in P'} \min_{g' \in G'} \|p' - g'\|\}. \end{aligned} \quad (9)$$

The 95% HD uses the 95th percentile of the distances between ground truth and prediction surface point sets. As a result, the impact of a very small subset of outliers is minimized when calculating HD.

4.3. Implementation Details

We implement UNETR in PyTorch[1] and MONAI[2]. The model was trained using a NVIDIA DGX-1 server. All models were trained with the batch size of 6, using the AdamW optimizer [31] with initial learning rate of 0.0001 for 20,000 iterations. For the specified batch size, the average training time

[1] http://pytorch.org/

[2] https://monai.io/

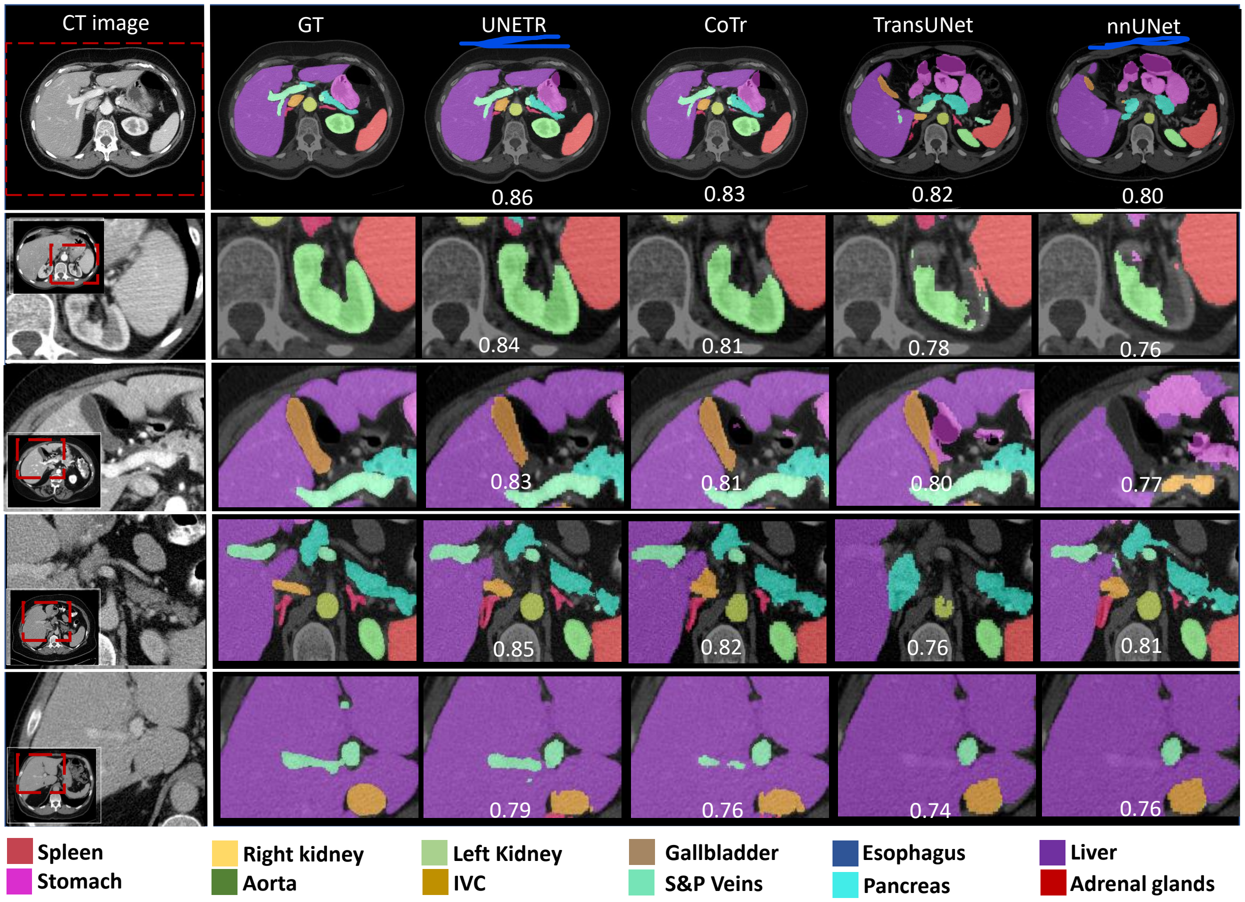


Figure 3. Qualitative comparison of different baselines in BTCV cross-validation. The first row shows a complete representative CT slice. We exhibit four zoomed-in subjects (row 2 to 5), where our method shows visual improvement on segmentation of kidney and spleen (row 2), pancreas and adrenal gland (row 3), gallbladder (row 4) and portal vein (row 5). The subject-wise average Dice score is shown on each sample.

was 10 hours for 20,000 iterations. Our transformer-based encoder follows the ViT-B16 [14] architecture with $L = 12$ layers, an embedding size of $K = 768$. We used a patch resolution of $16 \times 16 \times 16$. For inference, we used a sliding window approach with an overlap portion of 0.5 between the neighboring patches and the same resolution as specified in Sec. 4.1. We did not use any pre-trained weights for our transformer backbone (e.g. ViT on ImageNet) since it did not demonstrate any performance improvements. For BTCV dataset, we have evaluated our model and other baselines in the Standard and Free Competitions of its leaderboard. Additional data from the same cohort was used for the Free Competition increasing the number of training cases to 80 volumes. For all experiments, we employed five-fold cross validation with a ratio of 95:5. In addition, we used data augmentation strategies such as random rotation of 90, 180 and 270 degrees, random flip in axial, sagittal and coronal views and random scale and shift intensity. We used ensembling to fuse the outputs of models from four different five-fold cross-validations. For brain and spleen segmentation tasks in MSD dataset, we split the data into training, validation and test with a ratio of 80:15:5.

4.4. Quantitative Evaluations

UNETR outperforms the state-of-the-art methods for both Standard and Free Competitions on the BTCV leaderboard. As shown in Table 1, in the Free Competition, UNETR achieves an overall average Dice score of 0.899 and outperforms the second, third and fourth top-ranked methodologies by 1.238%, 1.696% and 5.269% respectively.

In the Standard Competition, we compared the performance of UNETR against CNN and transformer-based baselines. UNETR achieves a new state-of-the-art performance with an average Dice score of 85.3% on all organs. Specifically, on large organs, such as spleen, liver and stomach, our method outperforms the second best baselines by 1.043%, 0.830% and 2.125% respectively,in terms of Dice score. Furthermore, in segmentation of small organs, our method significantly outperforms the second best

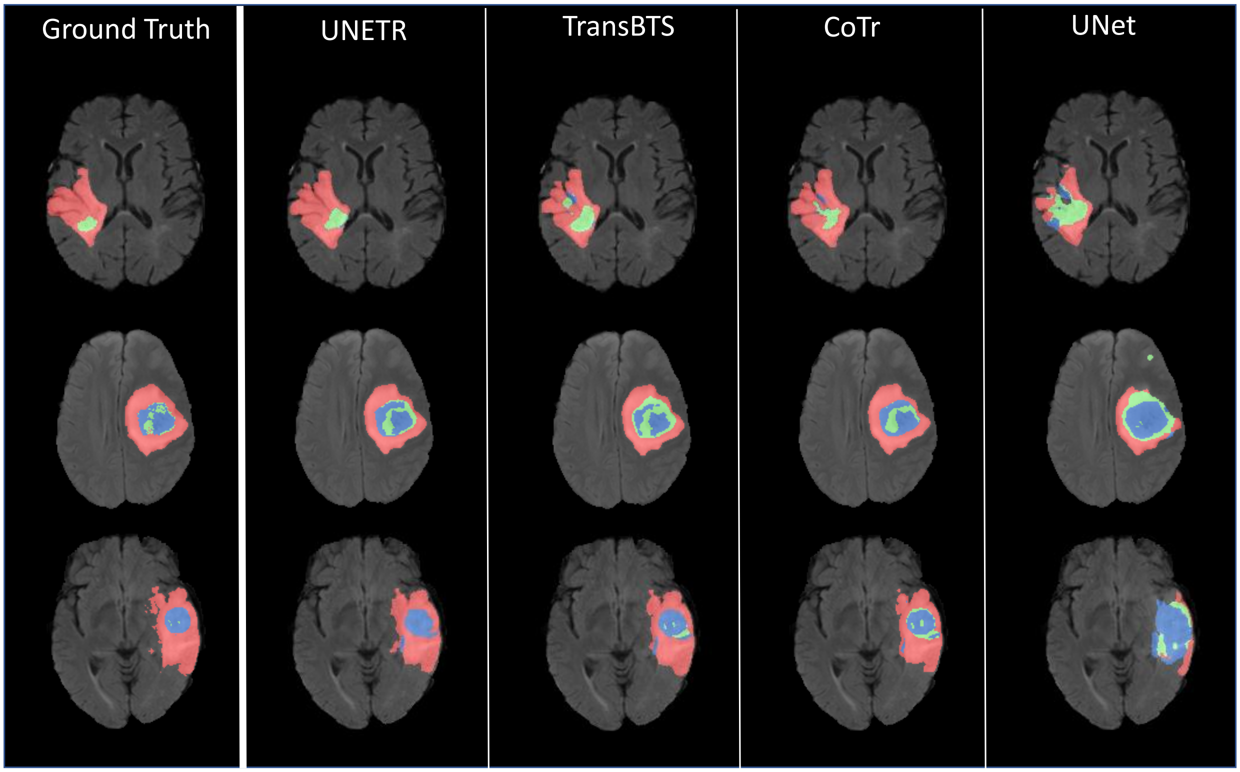


Figure 4. UNETR effectively captures the fine-grained details in segmentation outputs. The Whole Tumor (WT) encompasses a union of red, blue and green regions. The Tumor Core (TC) includes the union of red and blue regions. The Enhancing Tumor core (ET) denotes the green regions.

baselines by 6.382% and 6.772% on gallbladder and adrenal glands respectively, in terms of Dice score.

In Table 2, we compare the performance of UNETR against CNN and transformer-based methodologies for brain tumor and spleen segmentation tasks on MSD dataset. For brain segmentation, UNETR outperforms the closest baseline by 1.5% on average over all semantic classes. In particular, UNETR performs considerably better in segmenting tumor core (TC) subregion. Similarly for spleen segmentation, UNETR outperforms the best competing methodology by least 1.0% in terms of Dice score.

4.5. Qualitative Results

Qualitative multi-organ segmentation comparisons are presented in Fig. 3. UNETR shows improved segmentation performance for abdomen organs. Our model's capability of learning long-range dependencies is evident in row 3 (from the top), in which nnUNet confuses liver with stomach tissues, while UNETR successfully delineates the boundaries of these organs. In Fig. 3, rows 2 and 4 demonstrate a clear detection of kidney and adrenal glands against surrounding tissues, which indicate that UNETR captures better spatial context. In comparison to 2D transformer-based models, UNETR exhibits higher boundary segmentation accuracy as it accurately identifies the boundaries between kidney and spleen. This is evident for gallbladder in row 2, liver and stomach in row 3, and portal vein against liver in row 5. In Fig. 4, we present qualitative segmentation comparisons for brain tumor segmentation on the MSD dataset. Specifically, our model demonstrates better performance in capturing the fine-grained details of tumors.

5. Discussion

Our experiments in all datasets demonstrate superior performance of UNETR over both CNN and transformer-based segmentation models. Specifically, UNETR achieves better segmentation accuracy by capturing both global and local dependencies. In qualitative comparisons, this is illustrated in various cases in which UNETR effectively captures long-range dependencies (e.g. accurate segmentation of the pancreas tail in Fig. 3).

Moreover, the segmentation performance of UNETR on the BTCV leaderboard demonstrates new state-of-the-art benchmarks and validates its effectiveness. Specifically for small anatomies, UNETR outperforms both CNN and transformer-based models. Although 3D models already demonstrate high segmentation accuracy for small organs such as gallbladder, adrenal glands, UNETR can still outperform the best competing model by a significant margin (See Table 1). This is also observed in Fig. 3, in which

Organ	Spleen	Brain			
Decoder	Spleen	WT	ET	TC	All
NUP	0.932	0.721	0.527	0.660	0.636
PUP	0.941	0.749	0.558	0.698	0.668
MLA	0.950	0.757	0.563	0.732	0.684
UNETR	**0.964**	**0.789**	**0.585**	**0.761**	**0.711**

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

UNETR has a significantly better segmentation accuracy for left and right adrenal glands, and UNETR is the only model to correctly detect branches of the adrenal glands. For more challenging tissues, such as gallbladder in row 4 and portal vein in row 5, which have low contrast with the surrounding liver tissue, UNETR is still capable of segmenting clear connected boundaries.

6. Ablation

Decoder Choice In Table 3, we evaluate the effectiveness of our decoder by comparing the performance of UNETR with other decoder architectures on two representative segmentation tasks from MRI and CT modalities. In these experiments, we employ the encoder of UNETR but replaced the decoder with 3D counterparts of Naive UPsampling (NUP), Progressive UPsampling (PUP) and MuLti-scale Aggregation (MLA) [52]. We observe that these decoder architectures yield sub-optimal performance, despite MLA marginally outperforming both NUP and PUP. For brain tumor segmentation, UNETR outperforms its variants with MLA, PUP and NUP decoders by 2.7%, 4.3% and 7.5% on average Dice score. Similarly, for spleen segmentation, UNETR outerforms MLA, PUP and NUP by 1.4%, 2.3% and 3.2%.

Patch Resolution A lower input patch resolution leads to a higher sequence length, and therefore higher memory consumption, since it is inversely correlated to the cube of the resolution. As shown in Table 4, our experiments demonstrate that decreasing the resolution leads to consistently improved performance. Specifically, decreasing the patch resolution from 32 to 16 improves the performance by 1.1% and 0.8% in terms of average Dice score in spleen and brain segmentation tasks respectively. We did not experiment with lower resolutions due to memory constraints.

Organ	Spleen	Brain			
Resolution	Spleen	WT	ET	TC	All
32	0.953	0.776	0.579	0.756	0.703
16	**0.964**	**0.789**	**0.585**	**0.761**	**0.711**

Table 4. Effect of patch resolution on segmentation performance.

Models	#Params (M)	FLOPs (G)	Inference Time (s)
nnUNet [21]	19.07	412.65	10.28
CoTr [47]	46.51	399.21	19.21
TransUNet [7]	96.07	48.34	26.97
ASPP [11]	47.92	44.87	25.47
SETR [52]	86.03	43.49	24.86
UNETR	92.58	41.19	12.08

Table 5. Comparison of number of parameters, FLOPs and averaged inference time for various models in BTCV experiments.

Model and Computational Complexity In Table 5, we present number of FLOPs, parameters and averaged inference time of the models in BTCV benchmarks. Number of FLOPs and inference time are calculated based on an input size of $96 \times 96 \times 96$ and using a sliding window approach. According to our benchmarks, UNETR is a moderate-sized model with 92.58M parameters and 41.19G FLOPs. For comparison, other transformer-based methods such as CoTr [47], TransUNet [7] and SETR [52] have 46.51M, 96.07M and 86.03M parameters and 399.21G, 48.24G and 43.49G FLOPs, respectively. UNETR shows comparable model complexity while outperforming these models by a large margin in BTCV benchmarks. CNN-based segmentation models of nnUNet [21] and ASPP [10] have 19.07M and 47.92M parameters and 412.65G and 44.87G FLOPs, respectively. Similarly, UNETR outperforms these CNN-based models while having a moderate model complexity. In addition, UNETR has the second lowest averaged inference time after nnUNet [21] and is significantly faster than transformer-based models such as SETR [52], TransUNet [7] and CoTr [47].

7. Conclusion

This paper introduces a novel transformer-based architecture, dubbed as UNETR, for semantic segmentation of volumetric medical images by reformulating this task as a 1D sequence-to-sequence prediction problem. We proposed to use a 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 UNETR on different volumetric segmentation tasks in CT and MRI modalities. UNETR achieves new state-of-the-art performance in both Standard and Free Competitions on the BTCV leaderboard for the multi-organ segmentation and outperforms competing approaches for brain tumor and spleen segmentation on the MSD dataset. In conclusion, UNETR has shown the potential to effectively learn the critical anatomical relationships represented in medical images. The proposed method could be the foundation for a new class of transformer-based segmentation models in medical images analysis.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. 2016.

[2] Spyridon Bakas, Mauricio Reyes, et Int, and Bjoern Menze. 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.

[3] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

[4] Irwan Bello. Lambdanetworks: Modeling long-range interactions without attention. In *International Conference on Learning Representations*, 2020.

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294*, 2021.

[7] 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.

[8] Jianxu Chen, Lin Yang, Yizhe Zhang, Mark Alber, and Danny Z Chen. Combining fully convolutional and recurrent neural networks for 3d biomedical image segmentation. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 3044–3052, 2016.

[9] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. 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.

[10] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[11] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. *arXiv:1802.02611*, 2018.

[12] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting the design of spatial attention in vision transformers. *arXiv preprint arXiv:2104.13840*, 1(2):3, 2021.

[13] 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.

[14] 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 preprint arXiv:2010.11929*, 2020.

[15] 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 *International conference on medical image computing and computer-assisted intervention*, pages 149–157. Springer, 2016.

[16] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3146–3154, 2019.

[17] 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):1822–1834, 2018.

[18] Zaiwang Gu, Jun Cheng, Huazhu Fu, Kang Zhou, Huaying Hao, Yitian Zhao, Tianyang Zhang, Shenghua Gao, and Jiang Liu. Ce-net: Context encoder network for 2d medical image segmentation. *IEEE transactions on medical imaging*, 38(10):2281–2292, 2019.

[19] Nicholas Heller, Niranjan Sathianathen, Arveen Kalapara, Edward Walczak, Keenan Moore, Heather Kaluzniak, Joel Rosenberg, Paul Blake, Zachary Rengel, Makinna Oestreich, 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.

[20] Han Hu, Zheng Zhang, Zhenda Xie, and Stephen Lin. Local relation networks for image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3464–3473, 2019.

[21] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021.

[22] Fabian Isensee and Klaus H Maier-Hein. An attempt at beating the 3d u-net. *arXiv preprint arXiv:1908.02182*, 2019.

[23] Qiangguo Jin, Zhaopeng Meng, Changming Sun, Hui Cui, and Ran Su. Ra-unet: A hybrid deep attention-aware network to extract liver and tumor in ct scans. *Frontiers in Bioengineering and Biotechnology*, 8:1471, 2020.

[24] Konstantinos Kamnitsas, Liang Chen, Christian Ledig, Daniel Rueckert, and Ben Glocker. Multi-scale 3d convolutional neural networks for lesion segmentation in brain mri. *Ischemic stroke lesion segmentation*, 13:46, 2015.

[25] Konstantinos Kamnitsas, Christian Ledig, Virginia FJ Newcombe, Joanna P Simpson, Andrew D Kane, David K Menon, Daniel Rueckert, and Ben Glocker. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Medical image analysis*, 36:61–78, 2017.

[26] B Landman, Z Xu, J Igelsias, M Styner, T Langerak, and A Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, 2015.

[27] Wenqi Li, Guotai Wang, Lucas Fidon, Sebastien Ourselin, M Jorge Cardoso, and Tom Vercauteren. 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*, pages 348–360. Springer, 2017.

[28] Xiaomeng Li, Hao Chen, Xiaojuan Qi, Qi Dou, Chi-Wing Fu, and Pheng-Ann Heng. H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. *IEEE transactions on medical imaging*, 37(12):2663–2674, 2018.

[29] Siqi Liu, Daguang Xu, S Kevin Zhou, Olivier Pauly, Sasa Grbic, Thomas Mertelmeier, Julia Wicklein, Anna Jerebko, Weidong Cai, and Dorin Comaniciu. 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*, pages 851–858. Springer, 2018.

[30] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021.

[31] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[32] 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.

[33] Miguel Monteiro, Virginia FJ Newcombe, Francois Mathieu, Krishma Adatia, Konstantinos Kamnitsas, Enzo Ferrante, Tilak Das, Daniel Whitehouse, Daniel Rueckert, David K Menon, et al. Multiclass semantic segmentation and quantification of traumatic brain injury lesions on head ct using deep learning: an algorithm development and multicentre validation study. *The Lancet Digital Health*, 2(6):e314–e322, 2020.

[34] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

[35] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jonathon Shlens. Stand-alone self-attention in vision models. *arXiv preprint arXiv:1906.05909*, 2019.

[36] O. Ronneberger, P.Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proc. MICCAI*, volume 9351 of *LNCS*, pages 234–241, 2015.

[37] 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 preprint arXiv:1704.06382*, 2017.

[38] Amber L Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram Van Ginneken, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*, 2019.

[39] Yucheng Tang, Riqiang Gao, Ho Hin Lee, Shizhong Han, Yunqiang Chen, Dashan Gao, Vishwesh Nath, Camilo Bermudez, Michael R Savona, Richard G Abramson, et al. High-resolution 3d abdominal segmentation with random patch network fusion. *Medical Image Analysis*, 69:101894, 2021.

[40] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021.

[41] 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.

[42] 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.

[43] Wenxuan Wang, Chen Chen, Meng Ding, Jiangyun Li, Hong Yu, and Sen Zha. Transbts: Multimodal brain tumor segmentation using transformer. *arXiv preprint arXiv:2103.04430*, 2021.

[44] 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. *arXiv preprint arXiv:2102.12122*, 2021.

[45] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[46] Yingda Xia, Fengze Liu, Dong Yang, Jinzheng Cai, Lequan Yu, Zhuotun Zhu, Daguang Xu, Alan Yuille, and Holger Roth. 3d semi-supervised learning with uncertainty-aware multi-view co-training. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3646–3655, 2020.

[47] Yutong Xie, Jianpeng Zhang, Chunhua Shen, and Yong Xia. Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. *arXiv preprint arXiv:2103.03024*, 2021.

[48] Weijian Xu, Yifan Xu, Tyler Chang, and Zhuowen Tu. Co-scale conv-attentional image transformers. *arXiv preprint arXiv:2104.06399*, 2021.

[49] Lequan Yu, Xin Yang, Hao Chen, Jing Qin, and Pheng Ann Heng. Volumetric convnets with mixed residual connections for automated prostate segmentation from 3d mr images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[50] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International conference on machine learning*, pages 7354–7363. PMLR, 2019.

[51] Yundong Zhang, Huiye Liu, and Qiang Hu. Transfuse: Fusing transformers and cnns for medical image segmentation. *arXiv preprint arXiv:2102.08005*, 2021.

[52] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. *arXiv preprint arXiv:2012.15840*, 2020.

[53] Yuyin Zhou, Zhe Li, Song Bai, Chong Wang, Xinlei Chen, Mei Han, Elliot Fishman, and Alan L Yuille. Prior-aware neural

network for partially-supervised multi-organ segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10672–10681, 2019.

[54] 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)*, pages 178–184. IEEE, 2017.

[55] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.