IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 21, 2022, accepted February 24, 2022, date of publication March 4, 2022, date of current version March 21, 2022.

*Digital Object Identifier 10.1109/ACCESS.2022.3156894*

# Medical Image Segmentation Using Transformer Networks

**DAVOOD KARIMI[1], HAORAN DOU[2], AND ALI GHOLIPOUR[1], (Senior Member, IEEE)**

[1]Department of Radiology, Boston Children's Hospital, Harvard Medical School, Boston, MA 02115, USA
[2]Centre for Computational Imaging & Simulation Technologies in Biomedicine (CISTIB), School of Computing, University of Leeds, Leeds LS2 9JT, U.K.

Corresponding author: Davood Karimi (davood.karimi@childrens.harvard.edu)

This work was supported in part by the National Institute of Biomedical Imaging and Bioengineering and the National Institute of Neurological Disorders and Stroke of the National Institutes of Health (NIH) under Award R01EB031849, Award R01NS106030, and Award R01EB032366; in part by the Office of the Director of the NIH under Award S10OD0250111; and in part by a Technological Innovations in Neuroscience Award from the McKnight Foundation. The work of Haoran Dou was supported by the EPSRC Doctoral Training Partnership (DTP) Studentship. The content of this publication is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or the McKnight Foundation.

**ABSTRACT** Deep learning models represent the state of the art in medical image segmentation. Most of these models are fully-convolutional networks (FCNs), namely each layer processes the output of the preceding layer with convolution operations. The convolution operation enjoys several important properties such as sparse interactions, parameter sharing, and translation equivariance. Because of these properties, FCNs possess a strong and useful inductive bias for image modeling and analysis. However, they also have certain important shortcomings, such as performing a fixed and pre-determined operation on a test image regardless of its content and difficulty in modeling long-range interactions. In this work we show that a different deep neural network architecture, based entirely on self-attention between neighboring image patches and without any convolution operations, can achieve more accurate segmentations than FCNs. Our proposed model is based directly on the transformer network architecture. Given a 3D image block, our network divides it into non-overlapping 3D patches and computes a 1D embedding for each patch. The network predicts the segmentation map for the block based on the self-attention between these patch embeddings. Furthermore, in order to address the common problem of scarcity of labeled medical images, we propose methods for pre-training this model on large corpora of unlabeled images. Our experiments show that the proposed model can achieve segmentation accuracies that are better than several state of the art FCN architectures on two datasets. Our proposed network can be trained using only tens of labeled images. Moreover, with the proposed pre-training strategies, our network outperforms FCNs when labeled training data is small.

**INDEX TERMS** Deep learning, medical image segmentation, self-attention, transformer networks.

## I. INTRODUCTION

Image segmentation is needed for quantifying the size and shape of the volume/organ of interest, population studies, disease quantification, and computer-aided treatment and surgical planning. Given the importance and the difficulty of this task in medical applications, manual segmentation by a medical expert is regarded as the ground truth. However, manual segmentation is costly, time-consuming, and subject to inter and intra-observer disagreement. Automatic segmentation methods, on the other hand, have the potential to offer much faster, cheaper, and more reproducible results.

Classical techniques for medical image segmentation include region growing [1], deformable models [2], graph cuts [3], clustering methods [4], and Bayesian approaches [5]. Atlas-based methods are another very popular and powerful set of techniques [6]. With the introduction of deep learning methods for image segmentation [7], [8], these methods were quickly adopted for medical image segmentation. Deep learning methods have achieved unprecedented levels of performance on a range of medical image segmentation tasks [9]–[14]. One can argue that deep learning methods have largely replaced the classical methods for medical image segmentation.

Recent reviews of the main lines of research and recent advancements on the application of deep learning for medical image segmentation can be found in [15], [16]. Most recent studies have aimed at improving the network architecture, loss function, and training procedures. Recent works have shown that standard deep learning models can be trained using small numbers of labeled training images [17], [18]. Despite the large variability in the proposed network

The associate editor coordinating the review of this manuscript and approving it for publication was G. R. Sinha.

architectures, the one common feature in all of these works is that they all use the convolution operation as the main building block. The proposed architectures differ with regard to the arrangement of the convolutional operations, but they all rely on the same basic convolution operation. A few studies have proposed alternative network architectures based on recurrent neural networks [19], [20] and attention mechanisms [21]. There have also been attempts to improve the accuracy and robustness of these methods by modeling the statistical variation in the shape of the organ of interest and incorporating this shape information in the deep learning method [12], [22], [23]. However, all of those models still build upon the convolution operation. Some recent studies have suggested that a basic encoder-decoder-type fully convolutional network (FCN) can handle various segmentation tasks and be as accurate as more elaborate network architectures [24].

The convolution operation is also the main building block of the network architectures that have successfully addressed other central computer vision tasks such as image classification and object detection [25], [26]. These results attest to the effectiveness of the convolution operation for modeling and analyzing images. This effectiveness has been attributed to a number of key properties, including: 1) local (sparse) connections, 2) parameter sharing, and 3) translation equivariance [27], [28]. In fact, a convolutional layer can be regarded as a fully connected layer with an "infinitely strong prior" over its parameters [29].

The properties of the convolution operation that we mentioned above are, in part, inspired by neuroscience of the mammalian primary visual cortex [30]. They give convolutional neural networks (CNNs), including FCNs, a strong and useful inductive bias, which makes them highly effective and efficient in tackling different vision tasks. However, these same properties also put CNNs at some disadvantage. For example, the network weights are determined at training time and subsequently they are fixed. Therefore, these networks treat different images and different parts of an image equally. In other words, they lack a mechanism to change their weights depending on the image content. Furthermore, due to the local nature of convolution operations with small kernel sizes, CNNs cannot easily learn long-range interactions between distant parts of an image.

Attention-based neural network models have the potential to address some of the limitations of convolution-based models. In short, these models aim at learning the relationship between different parts of a sequence [31]. Most importantly, unlike CNNs, in attention-based networks not all network weights are fixed upon training. Rather, only a portion of the network weights are learned from training data and the rest of the weights are determined at test time based on the content of the input. Attention-based networks have become the dominant neural network architectures in natural language processing (NLP) applications. Transformers are the most common attention-based models in NLP [31]. Compared with recurrent neural networks, transformers can learn more complex and longer-range interactions much more effectively. Moreover, they overcome some of the central limitations of recurrent neural networks such as vanishing gradients. They also allow for parallel processing of inputs, which can lead to significantly shorter training time on modern hardware.

Despite the potential advantages of transformer networks, so far they have not been widely adopted in computer vision applications. A recent survey of the relevant works on this topic can be found in [32]. Application of attention-based neural networks for computer vision applications faces several important challenges. The number of pixels in a typical image is much larger than the length of a signal sequence (e.g., number of words) in typical NLP applications. This makes it impossible to directly apply standard attention models to images. The second main reason has been the training difficulty. The strong inductive bias of CNNs that we have mentioned above makes them highly data-efficient. Transformer networks, on the other hand, require much more training data because they incorporate minimal inductive bias. Recent studies have proposed practical solutions to these two challenges. To address the first challenge, vision transformer (ViT) proposed considering image patches, rather than pixels, as the units of information in an image [33]. ViT embeds image patches into a shared space and learns the relation between these embeddings using self-attention modules. It was shown that, given massive datasets of labeled images and vast computational resources, ViT could surpass CNNs in image classification accuracy. One possible solution to the second challenge was proposed in [34], where the authors used knowledge distillation from a CNN teacher to train a transformer network. It was shown that with this training strategy, transformer networks could achieve image classification accuracy levels on par with CNNs using the same amount of labeled training data [34].

In this work, we propose a self-attention-based deep neural network for 3D medical image segmentation. Our proposed network is based on self-attention between linear embeddings of 3D image patches, without any convolution operations. Given the fact that self-attention models generally require large labeled training datasets, we also propose unsupervised pre-training methods that can exploit large unlabeled medical image datasets. We compare our proposed model with several state of the art FCNs on two medical image segmentation datasets.

The specific contributions of this work are as follows:

1) We propose the first convolution-free deep neural network architecture for segmentation of 3D medical images.
2) We show that our proposed network can achieve segmentation performance levels that are better than or at least on par with the state of the art FCNs. Even though prior works have suggested that massive labeled training datasets are needed to effectively train transformer networks for NLP and vision applications, we experimentally show that our network can be trained using datasets of only $\sim 20-200$ labeled images.
3) We propose methods for pre-training our network on large corpora of unlabeled images. We show that when labeled training images are fewer in number, with these

pre-training strategies, our network performs better than a state of the art FCN with pre-training.

## II. MATERIALS AND METHODS

### A. PROPOSED NETWORK

Our proposed transformer network for 3D medical image segmentation is shown in Figure 1. The input to our network is a 3D image block $B \in \mathbb{R}^{W \times W \times W \times c}$, where $W$ denotes the extent of the block (in voxels) in each dimension and $c$ denotes the number of image channels. Working with image sub-blocks is a common approach in processing large volumetric images. It enables processing of large images of arbitrary size on limited GPU memory. Furthermore, it functions as an implicit data augmentation method because during training sub-blocks are sampled from random locations in the training images.

The image block $B$ is divided into $n^3$ non-overlapping 3D patches $\{p_i \in \mathbb{R}^{w \times w \times w \times c}\}_{i=1}^{N}$, where $w = W/n$ is the side length of each patch and $N = n^3$ denotes the number of patches in the block. In the experiments presented in this paper we choose $n \in \{3, 4, 5\}$, resulting in $N \in \{27, 64, 125\}$ patches in each block. The proposed transformer network embeds each patch into a lower-dimensional space and predicts the segmentation map corresponding to the image block $B$ based on the self-attention between these embeddings. The steps of the proposed method are described below.

Each of the $N$ patches $\{p_i\}_{i=1}^{N}$ is first reshaped into a vector of size $\mathbb{R}^{w^3 c}$ and embedded into $\mathbb{R}^D$ using a trainable linear mapping $E \in \mathbb{R}^{D \times w^3 c}$. This step is similar to the first step in the ViT model for image classification. The ViT model appended an extra "class token" to the sequence of embedded patches. This class token is inherited from NLP applications. We did not use such a token in the experiments presented in this work because our preliminary experiments showed that it did not improve the segmentation performance of our network in any way. Hence, the sequence of embedded patches $X^0 = [Ep_1; \ldots; Ep_N] + E_{\text{pos}}$ constitutes the input to our transformer network. The matrix $E_{\text{pos}} \in \mathbb{R}^{D \times N}$, which is added to the embedded patches is intended to learn a positional encoding. This is a common features of self-attention models because the attention mechanism is permutation-invariant. In other words, without such positional information, the transformer network ignores the ordering of the patches in the input sequence. In most NLP applications, the positional encoding has proved to be crucial for achieving optimal results. For 2D image classification with the ViT model, positional encoding resulted in relatively small improvements in performance and a simple 1D raster encoding was as good as more elaborate 2D positional encoding strategies [33]. Because we do not know a priori what type of positional encoding would be useful in the application considered in this work, we leave $E_{\text{pos}}$ as a free parameter to be learned along with the network parameters during training. In Section III, we present the results of experiments with different positional encoding strategies for our network.

As shown in Figure 1, our proposed network includes only the encoder section of the original transformer network proposed in [31]. The network has $K$ identical stages, each consisting of a multi-head self-attention (MSA) and a subsequent two-layer fully-connected feed-forward network (FFN). All MSA and FFN modules include residual connections, ReLU activations, and layer normalization [35]. Starting with the input sequence of embedded and position-encoded patches, $X^0$ described above, the $k^{\text{th}}$ stage of the network performs the following operations to map $X^k$ to $X^{k+1}$:

1) $X^k$ goes through $n_h$ independent heads in MSA. The $i^{\text{th}}$ head:
   a) Computes the query, key, and value sequences from the input sequence using linear operations:
   $$Q^{k,i} = E_Q^{k,i}\text{LN}(X^k), \quad K^{k,i} = E_K^{k,i}\text{LN}(X^k),$$
   $$V^{k,i} = E_V^{k,i}\text{LN}(X^k)$$
   where $E_Q, E_K, E_v \in \mathbb{R}^{D_h \times D}$ and LN denotes layer normalization.
   b) Computes the self-attention matrix and then the transformed values:
   $$A^{k,i} = \text{Softmax}(Q^T K)/\sqrt{D_h}$$
   $$\text{SA}^{k,i} = A^{k,i} V^{k,i}$$
   The above equation highlights one of the central differences between transformer networks and CNNs. It shows that the mapping ($A^{k,i}$) used to transform the features from one network layer to the next layer is computed based on the input itself. Hence, this mapping depends on the content of the input at test time, rather than being fixed and the same for all inputs as in CNNs.
2) Outputs of the $n_h$ self-attention heads are stacked together and re-projected back onto $\mathbb{R}^D$:
   $$\text{MSA}^k = E_{\text{reproj}}^k [\text{SA}^{k,0}; \ldots; \text{SA}^{k,n_h}]^T$$
   where $E_{\text{reproj}} \in \mathbb{R}^{D \times D_h n_h}$
3) The output of the current multi-head self-attention module is computed using a residual operation:
   $$X_{\text{MSA}}^k = \text{MSA}^k + X^k$$
4) $X_{\text{MSA}}^k$ goes through a two-layer FFN to obtain the output of the $k^{\text{th}}$ stage of the network:
   $$X^{k+1} = X_{\text{MSA}}^k + E_2^k\Big(\text{ReLU}\big(E_1^k\text{LN}(X_{\text{MSA}}^k) + b_1^k\big)\Big) + b_2^k$$

The output of the last stage, $X^K$, is passed through the final FFN layer that projects it onto $\mathbb{R}^{W^3 n_{\text{class}}}$. This is then reshaped into $\mathbb{R}^{W \times W \times W \times n_{\text{class}}}$. Here, $n_{\text{class}}$ denotes the number of classes (for binary segmentation, $n_{\text{class}} = 2$).

$$\hat{Y} = \text{Softmax}\Big(E_{\text{out}} X^K + b_{\text{out}}\Big).$$

$\hat{Y}$ is the predicted segmentation map for the block (as shown in Figure 1). Since our network predicts segmentation maps for image sub-blocks, in order to process a test image of arbitrary size, we apply the network in a sliding window fashion on the image.

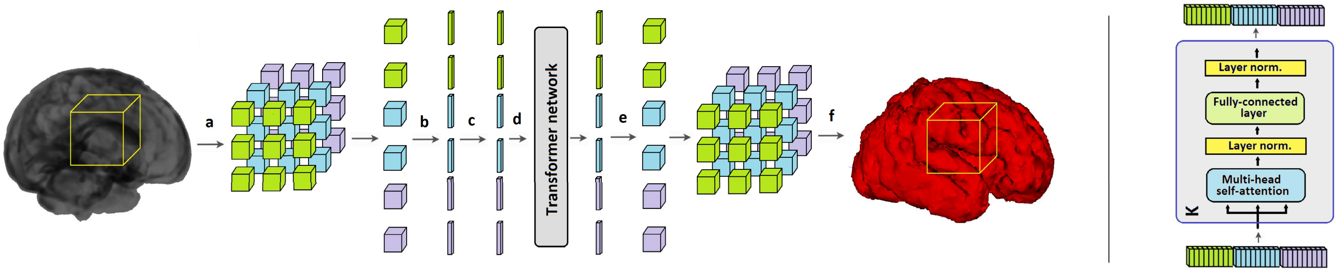


**FIGURE 1.** **The proposed convolution-free network for 3D medical image segmentation. Left: An overall schematic of the proposed method: (a) an image block is divided into $n^3$ non-overlapping 3D patches, (b) each patch is reshaped into a vector and embedded into a lower dimension, (c) positional encoding is added to the embedding, (d) position-encoded signals go through the transformer network, (e) the output of the network is re-projected back into the shape of the original patches, (f) the network output is the segmentation of the organ of interest for the location corresponding to the location of the extracted block. Right: One of the $K$ stages of the transformer network.**

### B. IMPLEMENTATION AND TRAINING

We implemented the network in TensorFlow 1.16 and trained it on an NVIDIA GeForce GTX 1080 GPU on a Linux machine with 120 GB of memory and 16 CPU cores. We compare our model with the following FCN architectures:

- 3D UNet++ [36]. This is a re-design of the UNet model [37]. The main difference between UNet++ and the basic UNet is a set of dense skip connections between the encoder and decoder sections of UNet++.
- Attention UNet [38]. This model is based on attention gates, which are meant to learn to automatically focus on the target organ. These attention gates enable the network to suppress irrelevant features and to learn useful soft region proposals, thereby improving segmentation performance.
- SE-FCN [39]. This network architecture is based on incorporating squeeze & excitation (SE) blocks [40] into FCNs for medical image segmentation. The purpose of SE blocks is to adaptively adjust the importance given to different feature maps, i.e., to promote more useful features and to down-weight less informative features.
- 3D Deeply Supervised Residual Network (DSRNet) [41]. This is an encoder-decoder FCN architecture that uses deep supervision [42] and skip connections between all corresponding encoder and decoder stages.

We trained the networks using a Dice similarity coefficient (DSC)-based loss function [43]:

$$\mathcal{L}(\hat{Y}, Y) = -\frac{\sum_i \hat{Y}_i Y_i}{\sum_i \hat{Y}_i^2 + \sum_i Y_i^2},$$

where Y is the ground truth segmentation map corresponding to the image block $B$ and the index $i$ runs over all voxels in the block. For training of our own network and the competing models we used the Adam optimization algorithm [44] with a batch size of 8. Furthermore, for all models we used blocks of size $24^3$ voxels. For our own network we used a learning rate of $10^{-4}$. For UNet++ a larger initial learning rate of $3 \times 10^{-4}$ was used because that led to the best results with UNet++. For Attention UNet, SE-FCN, and DSRNet we used a learning rate of $10^{-4}$.

### C. PRE-TRAINING

Manual segmentation of complex structures such as the brain cortical plate can take several hours of a medical expert's time for a single 3D image. Therefore, methods that can achieve high performance with fewer labeled training images are highly advantageous. This is especially important for transformer networks. As we mentioned above, transformer networks lack much of the built-in inductive bias that many other networks such as CNNs enjoy merely by the virtue of their architectural design. Therefore, compared with those architectures, transformers typically need much larger labeled training datasets in order to learn the underlying patterns directly from data. In NLP applications, a very common approach is to pre-train the network using unsupervised training on massive unlabeled datasets [45]. In the same spirit, we propose pretext tasks that can be used to train our network on unlabeled 3D medical image datasets.

#### 1) PRE-TRAINING WITH IMAGE DENOISING

In this approach, we add noise to the input image block $B$ and feed the noisy block $B_{\text{noisy}}$ to the network. We train the network to reconstruct the clean image block using an $\ell_2$ loss:

$$\mathcal{L}(B_{\text{noisy}}, B) = \|B_{\text{noisy}} - B\|_2.$$

The noise added to each voxel is independent and identically distributed Gaussian noise with SNR = 10 dB.

#### 2) PRE-TRAINING WITH IMAGE COMPLETION/INPAINTING

In this pre-training approach, we mask 10% of the image voxels at random. This is done by creating a random mask, $M \in \{0, 1\}^{W \times W \times W \times c}$, where each element of $M$ is a Bernoulli random variable with $p = 0.1$ and multiplying $M$ with $B$ in an element-wise fashion. The loss function used in this pre-training approach is similarly:

$$\mathcal{L}(B, M) = \|B - M \circ B\|_2.$$

For model pre-training with each of the above two strategies, we use a different output layer (without the softmax operation). In order to fine-tune the pre-trained network for the segmentation task, we introduce a new output layer with the softmax activation and train the network on the labeled data as explained above. We fine-tune the entire network, rather than only the output layer, on the labeled images because we have found that fine-tuning the entire network for the segmentation task leads to much better results.

Pre-training methods are also commonly used for FCNs. Prior studies have shown that pre-training might lead to substantial improvements in segmentation performance of FCNs, especially when the segmentation task is difficult and the size of labeled training data is small [17], [46]. Therefore, we will use the same denoising and inpainting tasks described above to pre-train the FCNs. Moreover, we will also use the semi-supervised FCN training method proposed in [47]. The method of [47] is based on an alternating optimization strategy. It alternately updates the network parameters and the estimated labels for the unlabeled images in parallel.

### *D. DATASETS AND EVALUATION CRITERIA*

Table 1 shows the datasets used for model training and evaluation in this work. The images were randomly split into training and test sets, with no patient data appearing in both training and test sets. The same training/test splits were used for all networks. For each dataset, we used approximately 20% of the training images for initial validation experiments to decide on training settings such as the learning rate for each network. After choosing the training settings, each network was trained on all training images. The only data augmentation was the implicit augmentation via sampling of image blocks from random locations in the training images. Voxel intensities of all images were normalized to have a zero mean and unit standard deviation. Moreover, all images were interpolated using 3D spline interpolation into isotropic voxel sizes shown in the table. The corresponding ground truth segmentations were interpolated using nearest neighbor interpolation. We compare our proposed method with the competing networks in terms of DSC, the 95 percentile of the Hausdorff Distance (HD95), and Average Symmetric Surface Distance (ASSD).

## III. RESULTS AND DISCUSSION

Table 2 compares the segmentation performance of the proposed method with the competing FCNs on the brain cortical plate and hippocampus datasets. As described in Section II-A, the proposed network includes several hyper-parameters that can influence the segmentation results. The results presented in Table 2 were obtained with: $K = 5, W = 24, n = 3, D = 1024, D_h = 256, n_h = 4$. These are our default settings for network hyper-parameters that we have used in all experiments reported in the rest of the paper, unless otherwise specified. We arrived at these parameters using cross-validation experiments on the training images in the brain cortical plate and hippocampus datasets as well as other datasets not presented in the paper. We present experimental results on the effects of different hyper-parameters on the segmentation performance below.

The results presented in Table 2 show that the proposed network has achieved segmentation performance levels that are superior to the competing FCNs. For each dataset and each of the three criteria, we performed paired t-tests to see if the differences were statistically significant. As shown in the table, segmentation performance of the proposed convolution-free network was significantly better than the four FCNs in terms of DSC, HD95, and ASSD at $p < 0.01$. Specifically, for paired t-tests between the proposed model and UNet++ on the brain cortical plate dataset the p-values for DSC, HD95, and ASSD were, respectively, 0.0044, $31 \times 10^{-5}$, and 0.0082. For the hippocampus dataset, the p-values for these comparisons were, respectively, $20 \times 10^{-6}$, 0.0032, and $74 \times 10^{-6}$. The results obtained with the proposed method were especially superior in terms of the distance metrics, i.e., HD95 and ASSD. Among the FCN architectures, UNet++ performed substantially better than the other architectures on both datasets, but its segmentation performance was significantly inferior to that of our proposed method.

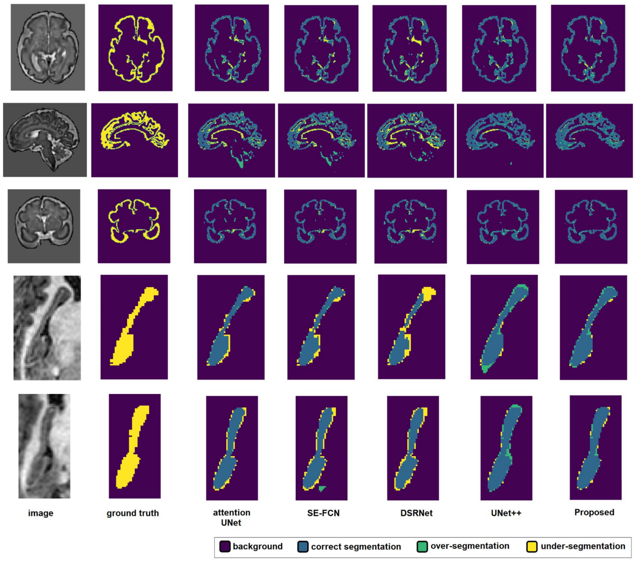


**FIGURE 2. Example segmentations predicted by the proposed method and the four FCNs. The segmentation legend is shown at the top of the figure. Three example slices from the brain cortical plate dataset and two example slices from the hippocampus dataset are shown. In each example, the first row shows the image slice and the ground-truth segmentation map. The second row shows the predictions of the four FCNs and the proposed transformer network.**

Figure 2 shows example slices from test images in each dataset and the segmentations predicted by the proposed method and the four FCNs. Visual inspection of the results shows that the proposed network is capable of accurately segmenting fine and intricate structures such as the brain cortical plate. On both datasets, Attention UNet, DSRnet, and SEFCN often resulted in false positive predictions far away from the target organ, which is the reason behind their poor performance in terms of the distance metrics presented in Table 2.

We further assessed the segmentation performance of our proposed network with reduced number of labeled training images. The goal of this experiment was to investigate if the pre-training tasks proposed in Section II-C could help the network achieve a good segmentation performance with

**TABLE 1.** Datasets used for experiments in this work.

| target organ | image modality | $[n_{train}, n_{test}]$ | image resolution (mm) | source |
|---|---|---|---|---|
| Brain cortical plate | T2 MRI | [18, 9] | 0.80 | In-house (Boston Children's Hospital) |
| Hippocampus | MRI | [220, 40] | 0.75 | https://decathlon-10.grand-challenge.org/ |

**TABLE 2.** Comparison of the segmentation performance of the proposed method and several competing FCNs on the brain cortical plate and hippocampus datasets. Better results for each dataset/criterion have been marked using bold type. We used paired t-tests to find statistically significant differences; asterisks denote significantly better results at $p < 0.01$.)

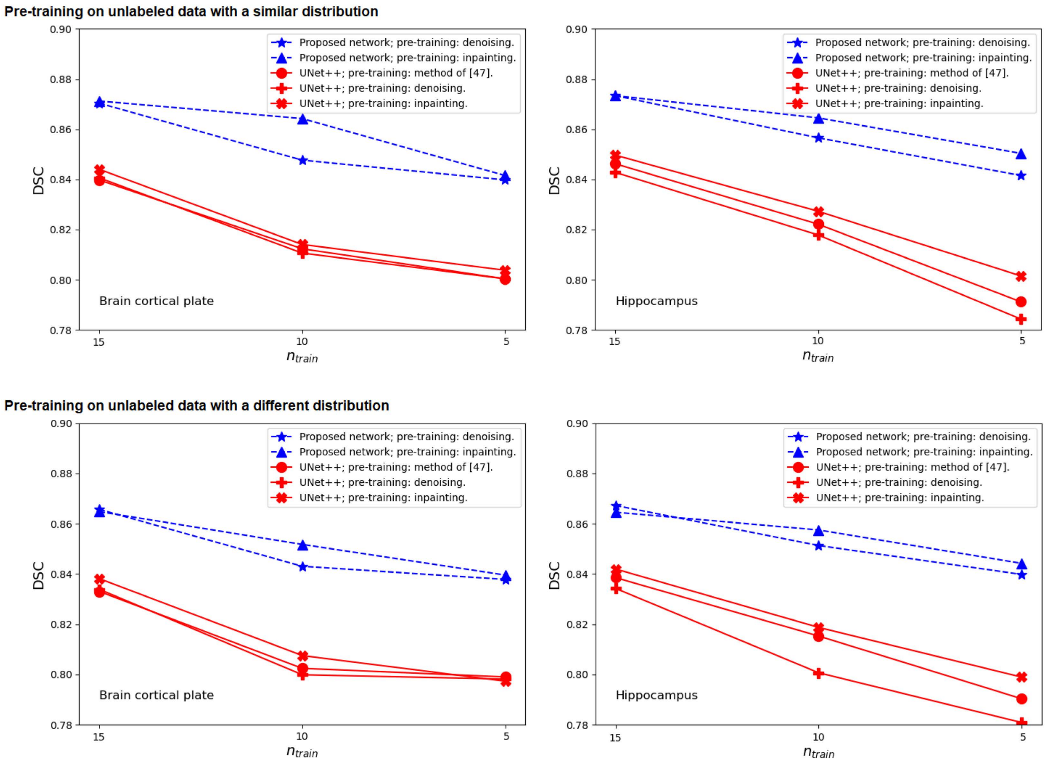
| Dataset | Method | DSC | HD95 (mm) | ASSD (mm) |
|---|---|---|---|---|
| Brain cortical plate | Proposed | $\mathbf{0.878 \pm 0.037}^*$ | $\mathbf{0.871 \pm 0.141}^*$ | $\mathbf{0.271 \pm 0.068}^*$ |
| | UNet++ | $0.862 \pm 0.054$ | $0.938 \pm 0.430$ | $0.258 \pm 0.082$ |
| | Attention UNet | $0.831 \pm 0.060$ | $1.125 \pm 0.337$ | $0.536 \pm 0.128$ |
| | DSRnet | $0.846 \pm 0.050$ | $0.978 \pm 0.229$ | $0.492 \pm 0.135$ |
| | SEFCN | $0.786 \pm 0.076$ | $2.095 \pm 0.801$ | $0.641 \pm 0.115$ |
| Hippocampus | Proposed | $\mathbf{0.895 \pm 0.020}^*$ | $\mathbf{1.035 \pm 0.203}^*$ | $\mathbf{0.416 \pm 0.064}^*$ |
| | UNet++ | $0.874 \pm 0.027$ | $1.479 \pm 1.427$ | $0.533 \pm 0.208$ |
| | Attention UNet | $0.820 \pm 0.036$ | $5.850 \pm 5.150$ | $1.192 \pm 0.597$ |
| | DSRnet | $0.821 \pm 0.038$ | $3.438 \pm 3.065$ | $1.395 \pm 0.316$ |
| | SEFCN | $0.716 \pm 0.052$ | $8.614 \pm 5.230$ | $1.795 \pm 0.761$ |

a small number of labeled training images. In this experiment, we compared our model with UNet++, which was more accurate than the other three FCNs in the experiments presented in Table 2. For this experiment, we trained our method and UNet++ using $n_{\text{train}} = 5, 10,$ and 15 labeled training images from cortical plate and hippocampus datasets. We pre-trained our network using either the denoising or the in-painting tasks described in Section II-C. We pre-trained UNet++ using the same denoising and inpainting pre-training tasks and also using the method proposed in [47]. Furthermore, we performed this experiment in two different ways:

1) **Pre-training on data with a similar distribution.** For brain cortical plate segmentation, we used 500 T2 brain images from the developing Human Connectome Project (dHCP) dataset [14] for pre-training. The subjects in the dHCP dataset range in age between 29 and 44 gestational weeks, which is close enough to the age range of our in-house dataset: between 16 and 39 gestational weeks. For hippocampus segmentation, we used the remaining training images (i.e., $220-n_{\text{train}}$) for pre-training.
2) **Pre-training on data with a different distribution.** Sometimes even unlabeled images with the same distribution are not available. To simulate such a scenario, we used a pool of publicly available computed tomography (CT) images. Specifically, we used 130 liver CT [48] and 300 kidney CT [49] images to pre-train our network and UNet++ for both brain cortical plate segmentation and hippocampus segmentation. As we had done for our target MRI images described above, we also normalized voxel intensities of these CT images to have a zero mean and unit standard deviation.

Figure 3 shows the results of this experiment. The results show that with the proposed pre-training, our convolution-free network achieves significantly more accurate segmentations with fewer labeled training images. As expected, on both datasets there was a drop in the segmentation performance as the number of labeled training images was reduced. However, this drop was smaller for the proposed network than for UNet++. We have observed very similar results with other FCN architectures. For our network as well as for UNet++, the proposed inpainting pre-training leads to slightly better results than the other pre-training methods. Moreover, overall, pre-training on similar images leads to better segmentation performance than pre-training on a dataset of different images. As shown in Figure 3, for both the proposed network and UNet++ the segmentation performance is, slightly but consistently, higher when pre-training is performed on a similar dataset. This indicates that both the proposed network and UNet++ can learn the existing patterns in unlabeled images and leverage this information to achieve better segmentation results.

This is a very interesting and promising observation because it shows that the proposed network can be trained using a handful of labeled images for segmenting complex structures in 3D medical images. This result is even more noteworthy when we consider the results reported by recent image classification studies. As we explained in Section I, image classification studies that used a similar approach (i.e., applying a transformer network on patch embeddings) required massive labeled datasets [33] or relied on knowledge distillation from a CNN teacher model [34]. Our results, on the other hand, show that only a handful of labeled training images are sufficient to train a similar network for 3D medical image segmentation. This can be attributed to several factors: 1) In image classification there are significant variations in relevant image features (even among images that belong to the same class). In the segmentation tasks considered here, on the other hand, there is significant similarity across subjects and even among different patches in the same image. 2) There are far fewer class labels (only two) in the segmentation tasks considered here compared with image classification applications. 3) Working with image sub-blocks acts


Pre-training on unlabeled data with a similar distribution

Pre-training on unlabeled data with a different distribution

**FIGURE 3. Segmentation performance (in terms of DSC) for the proposed network and UNet++ with reduced number of labeled training images on the brain cortical plate dataset (left) and the hippocampus dataset (right). The top two plots are for the experiment where the images used for pre-training have a distribution that is similar to the target test images. The bottom two plots are for the experiment where the images used for pre-training (liver and kidney CT images) are very different from the target test images.**

as a strong data augmentation strategy and enables optimal utilization of labeled training images. As a result, despite their minimal inductive bias, transformer networks appear to be well suited for medical image segmentation tasks.

The experimental results presented above show that the proposed method can achieve segmentation performance on par with or better than FCNs with as few as 10-20 labeled training images. This is an important and encouraging result because in the medical imaging domain manual labels are not easy to obtain. Nonetheless, in order to assess the performance of the proposed method on larger datasets, we conducted another experiment with the newborn brain scans in the developing Human Connectome Project (dHCP) dataset [14]. This dataset includes 558 T2 MRI brain scans with cortical plate segmentation. We randomly selected 58 of these scans as test images. We then trained our model and UNet++ on all 500 remaining images as well as subsets of 100 and 10 images. In addition to the implicit data augmentation caused by sampling patches from random locations in the training images, we applied random flip and rotation and we added random Gaussian noise to the images. We also experimented with random down/up-scaling of the images and random elastic deformation, but these augmentations had a negative impact on segmentation performance because they reduced the accuracy of training labels for fine and complex cortical plate segmentation. Therefore, we did not use these latter augmentation methods. The results of this experiment are presented in Table 3. The results indicate that the proposed method achieves better segmentation results than UNet++ with either 10, 100, or 500 labeled training images. Paired t-tests on the 58 test images showed that, with 500 labeled training images, the proposed method achieved significantly higher DSC ($p = 0.0015$) and significantly lower ASSD ($p = 84 \times 10^{-6}$).

In Figures 5 and 4, we have shown example attention maps of the proposed network for two different datasets. As mentioned above, in order to process a test image of arbitrary size, we apply our network in a sliding window fashion. To generate the attention maps for the whole image, at each location of the sliding window the attention matrices (which are of size $\mathrm{I\!R}^{N\times N}$) are summed along their columns to determine the total attention paid to each of the $N$ patches by the other patches in the block. Performing this computation in a sliding window fashion and computing the voxel-wise average gives us the attention maps shown in these figures. They indicate how much attention is paid to every part of the image in the process of generating the segmentation map.

The attention maps shown in Figure 4 were generated on pancreas CT images from the Medical Segmentation Decathlon challenge (https://decathlon-10.grand-challenge.org/). The attention maps show that, overall, the early stages of the network have a wider attention scope.

**TABLE 3. Results of an experiment to investigate the performance of the proposed network with different numbers of training images. This experiment was performed on brain cortical plate segmentation labels from the dHCP dataset [14]. Asterisks denote statistically significant differences at $p = 0.01$, computed using paired t-tests.**

| No. of training images | Method | DSC | HD95 (mm) | ASSD (mm) |
|---|---|---|---|---|
| $n_{\text{train}} = 10$ | Proposed | $\mathbf{0.910 \pm 0.033}$* | $\mathbf{0.814 \pm 0.042}$ | $\mathbf{0.231 \pm 0.070}$* |
| | UNet++ | $0.884 \pm 0.040$ | $0.816 \pm 0.043$ | $0.247 \pm 0.065$ |
| $n_{\text{train}} = 100$ | Proposed | $\mathbf{0.926 \pm 0.030}$* | $\mathbf{0.806 \pm 0.040}$* | $\mathbf{0.226 \pm 0.068}$* |
| | UNet++ | $0.916 \pm 0.032$ | $0.813 \pm 0.044$ | $0.241 \pm 0.062$ |
| $n_{\text{train}} = 500$ | Proposed | $\mathbf{0.928 \pm 0.031}$* | $0.805 \pm 0.041$ | $\mathbf{0.221 \pm 0.070}$* |
| | UNet++ | $0.916 \pm 0.028$ | $0.805 \pm 0.044$ | $0.240 \pm 0.057$ |

**TABLE 4. Effect of some of the network hyperparameters on the segmentation performance on the brain cortical plate dataset. The baseline network (shown in the first row of this table) corresponds to these settings: $K = 5$, $W = 24$, $n = 3$, $D = 1024$, $D_h = 256$, $n_h = 4$, which are the hyperparameter values that we used in all experiments reported in this paper other than in this table. In the remaining rows of this table, we have changed the hyperparameters, one at a time, and trained the network.**

| Parameter settings | DSC | HD95 (mm) | ASSD (mm) |
|---|---|---|---|
| baseline | $0.878 \pm 0.037$ | $0.871 \pm 0.141$ | $0.271 \pm 0.068$ |
| larger blocks, $n = 5$ | $0.881 \pm 0.033$ | $0.821 \pm 0.113$ | $0.257 \pm 0.058$ |
| larger patches, $W = 36$ | $0.865 \pm 0.043$ | $0.898 \pm 0.155$ | $0.286 \pm 0.069$ |
| no positional encoding | $0.863 \pm 0.056$ | $0.904 \pm 0.164$ | $0.296 \pm 0.077$ |
| fixed positional encoding | $0.866 \pm 0.050$ | $0.881 \pm 0.144$ | $0.274 \pm 0.060$ |
| deeper network, $K = 10$ | $0.857 \pm 0.070$ | $0.980 \pm 0.171$ | $0.301 \pm 0.080$ |
| shallower network, $K = 3$ | $0.869 \pm 0.081$ | $1.030 \pm 0.176$ | $0.297 \pm 0.086$ |
| more heads, $n_h = 8$ | $0.877 \pm 0.031$ | $0.864 \pm 0.147$ | $0.264 \pm 0.060$ |
| single head, $n_h = 1$ | $0.845 \pm 0.048$ | $1.112 \pm 0.208$ | $0.325 \pm 0.084$ |

They attend to other structures and anatomical features that surround the organ of interest (here, the pancreas). The deeper stages of the network are more focused on the pancreas itself. A rather similar pattern can be observed in the segmentation maps for brain cortical plate segmentation that are shown in Figure 5. In the earlier stages, the network attends to the entire brain, while in the deeper layers the network's attention tends to be more focused to the regions around the cortical plate.

Another observation from these figures, especially Figure 4, is the variability among the attention patterns of different heads in a multi-head self-attention (MSA) module. In each stage, the four separate heads of the MSA module adopt quite different attention strategies. This may suggest that the multi-head design of the MSA module gives the network more flexibility, enabling it to learn more complex attention patterns that help improve the segmentation performance. The importance of multi-head design is well documented in natural language processing applications [31], and our results show that it is important for 3D medical image segmentation as well. We further show this below by quantifying the effect of the number of attention heads on segmentation performance.

Table 4 shows the results of a set of experiments on the brain cortical plate dataset to investigate the effects of some of the network hyper-parameters on segmentation performance. In this table, the baseline network (first row) corresponds to the settings that we have used in the experiments reported above, i.e., $K = 5$, $W = 24$, $n = 3$, $D = 1024$, $D_h = 256$, and $n_h = 4$. We chose these settings based on preliminary experiments and we have found them to work well for different datasets.

The results presented in this table show that, overall, the performance of the network is not very sensitive to the hyper-parameter settings. For example, changing the number or the size of the patches typically leads to slight changes in performance. We have also observed that a network depth of $K \in [5, 7]$ leads to best results, whereas much deeper or shallower networks were not better. Furthermore, using a fixed positional encoding or no positional encoding slightly reduces the segmentation performance compared with free-parameter/learnable positional encoding. Finally, using a single-head attention significantly reduces the segmentation performance, which indicates the importance of the multi-head design to enable the network to learn a more complex relation between neighboring patches.

Many of the above observations are consistent with the experimental results that have been reported in other applications. For example, our results show that increasing the number of MSA heads ($n_h$) or the network depth ($K$) beyond a certain limit has a negative impact on segmentation performance. This observation is similar to some of the experimental results reported in [33], [50], where networks with a larger number of MSA heads and/or larger number of layers resulted in lower image classification accuracy on several datasets. Similar results have been reported in natural language processing applications [51]. For example, one study showed that it was possible to prune 50-72% of the attention heads without a significant reduction in model accuracy in a machine translation application [52]. This is because, depending on the application, a certain number of heads are sufficient to learn the attention patterns between the signals in a sequence (i.e., patch embeddings in our application). Further increasing the number of heads will only increase the number of network parameters without providing any useful capacity to the network.

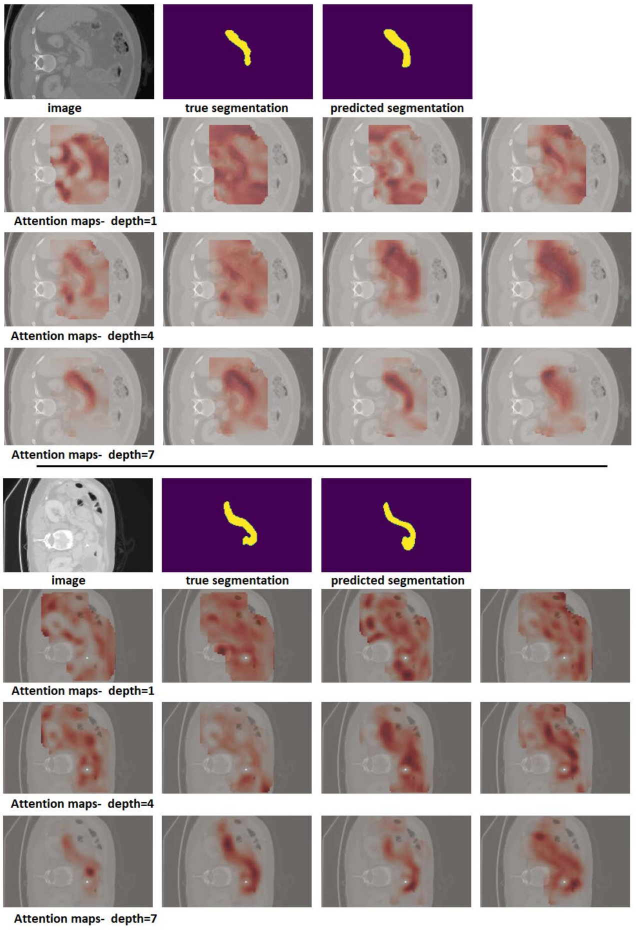


**FIGURE 4.** Example attention maps for two pancreas images. In this experiment, a network with a depth $K = 7$ was used. Attention maps for depths 1, 4, and 7 are shown. Each row shows the attention map for one of the four heads.

There may be other factors that can influence the relative advantages of the proposed transformer network compared with FCNs. Some of these factors are image resolution, size of the organ/volume of interest, and patch size. Our experiments show that these factors do not affect the superiority of the proposed method over FCNs. For example, we resampled the brain cortical plate and hippocampus datasets to isotropic voxel sizes of 0.5 mm and repeated our experiments. The results showed that the proposed network achieved significantly higher DSC and significantly lower HD95 and ASSD than UNet++ on both datasets. We also applied UNet++ on larger patch sizes of $48^3$ and $64^3$ voxels. This change did not improve the performance of UNet++ on the brain cortical plate dataset. It slightly improved the segmentation performance of UNet++ on the hippocampus dataset (DSC: $0.877\pm0.029$, HD95: $1.448\pm1.430$ mm, and ASSD: $0.502\pm 0.201$ mm). However, these were still statistically inferior to those obtained with our proposed network (Table 2). Increasing the input image block size to $48^3$ or $64^3$ voxels did not significantly improve the performance of our proposed network either.

Table 5 shows the number of learnable parameters, number of floating point operations (FLOPS), and frames per second (FPS). FLOPS and FPS are reported for processing patches of size $24^3$ voxels. We computed the FPS for all models on an NVIDIA RTX 2080TI GPU. Overall, the models have relatively similar number of parameters and computational costs. Our proposed network has a slightly smaller number of parameters than the compared FCNs. On the other hand, the number of FLOPS for the proposed network is higher, which is due to the large matrix multiplications involved in the attention modules. In terms of training time, our network converged in approximately 24 hours of GPU time, whereas the FCNs converged in approximately 4 hours of training. This might be due to the fact that transformer networks need additional training time in order to internalize the spatial patterns in the image, whereas FCNs' architecture makes this learning easier.

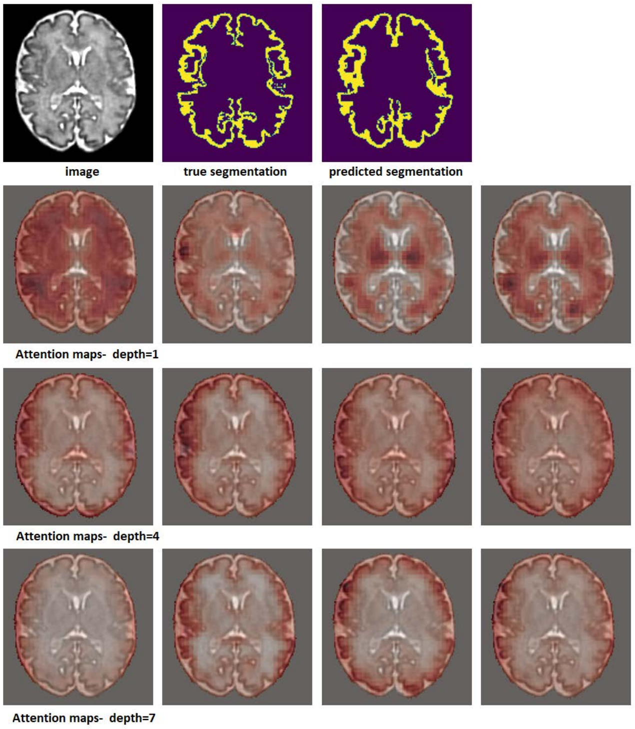


**FIGURE 5.** Example attention maps for a cortical plate image. In this experiment, a network with a depth $K = 7$ was used. Attention maps for depths 1, 4, and 7 are shown. Each row shows the attention map for one of the four heads.

**TABLE 5.** The number of free parameters ($n_{\text{param}}$), number of floating point operations (FLOPS), and frames per second (FPS) for each of the FCNs and the proposed network. FLOPS and FPS are computed for processing patches of size $24^3$ voxels.

| Model | $n_{\text{param}} \times 10^6$ | FLOPS $\times 10^9$ | FPS |
|---|---|---|---|
| SEFCN | 2.60 | 4.16 | 190.3 |
| DSRNet | 3.70 | 3.52 | 173.2 |
| Attention UNet | 2.77 | 1.83 | 157.0 |
| UNet++ | 3.72 | 4.02 | 160.2 |
| Proposed | 2.49 | 4.88 | 144.5 |

## IV. CONCLUSION

The convolution operation has a strong basis in the structure of the mammalian primary visual cortex and it is well suited for developing powerful techniques for image modeling and image understanding. In recent years, CNNs have been shown to be highly effective in tackling various computer vision problems. However, there is no reason to expect that no other model can outperform CNNs on a specific vision task. Medical image analysis applications, in particular, pose specific challenges such as 3D nature of the images and small number of labeled images. In such applications, other models could be more effective than CNNs. In this work we presented a new model for 3D medical image segmentation. Unlike all recent models that use convolutions as their main building blocks, our model is based on self-attention between neighboring 3D patches. Our results show that the proposed network can outperform state of the art FCNs on three medical image segmentation datasets. With pre-training for denoising and in-painting tasks on unlabeled images, our network also performed better than an FCN when only 5-15 labeled training images were available. We expect that the network proposed in this paper should be effective for other tasks in medical image analysis such as anomaly detection and classification.

## REFERENCES

[1] P. Gibbs, D. L. Buckley, S. J. Blackband, and A. Horsman, "Tumour volume determination from MR images by morphological segmentation," *Phys. Med. Biol.*, vol. 41, no. 11, p. 2437, 1996.

[2] Y. Wang, Q. Guo, and Y. Zhu, "Medical image segmentation based on deformable models and its applications," in *Deformable Models*. New York, NY, USA: Springer, 2007, pp. 209–260.

[3] D. Mahapatra and J. M. Buhmann, "Prostate MRI segmentation using learned semantic knowledge and graph cuts," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 3, pp. 756–764, Mar. 2014.

[4] A. F. Goldszal, C. Davatzikos, D. L. Pham, M. X. H. Yan, R. N. Bryan, and S. M. Resnick, "An image-processing system for qualitative and quantitative volumetric analysis of brain images," *J. Comput. Assist. Tomogr.*, vol. 22, no. 5, pp. 827–837, Sep. 1998.

[5] J. L. Prince, D. Pham, and Q. Tan, "Optimization of MR pulse sequences for Bayesian image segmentation," *Med. Phys.*, vol. 22, no. 10, pp. 1651–1656, Oct. 1995.

[6] P. M. Thompson and A. W. Toga, "Detection, visualization and animation of abnormal anatomic structure with a deformable probabilistic brain atlas based on random vector field transformations," *Med. Image Anal.*, vol. 1, no. 4, pp. 271–294, Sep. 1997.

[7] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[8] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[9] S. Bakas *et al.*, "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge," 2018, *arXiv:1811.02629*.

[10] O. Bernard *et al.*, "Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved?" *IEEE Trans. Med. Imag.*, vol. 37, no. 11, pp. 2514–2525, Nov. 2018.

[11] K. Kamnitsas, C. Ledig, V. F. J. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker, "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation," *Med. Image Anal.*, vol. 36, pp. 61–78, Feb. 2017.

[12] D. Karimi, Q. Zeng, P. Mathur, A. Avinash, S. Mahdavi, I. Spadinger, P. Abolmaesumi, and S. E. Salcudean, "Accurate and robust deep learning-based segmentation of the prostate clinical target volume in ultrasound images," *Med. Image Anal.*, vol. 57, pp. 186–196, Oct. 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1361841519300623

[13] Q. Zeng, D. Karimi, E. H. T. Pang, S. Mohammed, C. Schneider, M. Honarvar, and S. E. Salcudean, "Liver segmentation in magnetic resonance imaging via mean shape fitting with fully convolutional neural networks," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2019, pp. 246–254.

[14] M. Bastiani, J. L. R. Andersson, L. Cordero-Grande, M. Murgasova, J. Hutter, A. N. Price, A. Makropoulos, S. P. Fitzgibbon, E. Hughes, D. Rueckert, S. Victor, M. Rutherford, A. D. Edwards, S. M. Smith, J.-D. Tournier, J. V. Hajnal, S. Jbabdi, and S. N. Sotiropoulos, "Automated processing pipeline for neonatal diffusion MRI in the developing human connectome project," *NeuroImage*, vol. 185, pp. 750–763, Jan. 2019.

[15] M. H. Hesamian, W. Jia, X. He, and P. Kennedy, "Deep learning techniques for medical image segmentation: Achievements and challenges," *J. Digit. Imag.*, vol. 32, no. 4, pp. 582–596, Aug. 2019.

[16] S. A. Taghanaki, K. Abhishek, J. P. Cohen, J. Cohen-Adad, and G. Hamarneh, "Deep semantic segmentation of natural and medical images: A review," *Artif. Intell. Rev.*, vol. 54, no. 1, pp. 1–42, 2020.

[17] M. Ghafoorian, A. Mehrtash, T. Kapur, N. Karssemeijer, E. Marchiori, M. Pesteie, C. R. G. Guttmann, F.-E. de Leeuw, C. M. Tempany, B. van Ginneken, A. Fedorov, P. Abolmaesumi, B. Platel, and W. M. Wells, III, "Transfer learning for domain adaptation in MRI: Application in brain lesion segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2017, pp. 516–524.

[18] D. Karimi, S. K. Warfield, and A. Gholipour, "Critical assessment of transfer learning for medical image segmentation with fully convolutional neural networks," 2020, *arXiv:2006.00356*.

[19] Y. Gao, J. M. Phillips, Y. Zheng, R. Min, P. T. Fletcher, and G. Gerig, "Fully convolutional structured LSTM networks for joint 4D medical image segmentation," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 1104–1108.

[20] W. Bai, H. Suzuki, C. Qin, G. Tarroni, O. Oktay, P. M. Matthews, and D. Rueckert, "Recurrent neural networks for aortic image sequence segmentation with sparse annotations," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2018, pp. 586–594.

[21] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.

[22] O. Oktay, E. Ferrante, K. Kamnitsas, M. Heinrich, W. Bai, J. Caballero, S. A. Cook, A. De Marvao, T. Dawes, D. P. O'Regan, and B. Kainz, "Anatomically constrained neural networks (ACNNs): Application to cardiac image enhancement and segmentation," *IEEE Trans. Med. Imag.*, vol. 37, no. 2, pp. 384–395, Feb. 2017.

[23] D. Karimi, G. Samei, C. Kesch, G. Nir, and S. E. Salcudean, "Prostate segmentation in MRI using a convolutional neural network architecture and training strategy based on statistical shape models," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 13, no. 8, pp. 1211–1219, Aug. 2018, doi: 10.1007/s11548-018-1785-8.

[24] F. Isensee, P. Kickingereder, W. Wick, M. Bendszus, and K. H. Maier-Hein, "No new-net," in *Proc. Int. MICCAI Brainlesion Workshop*. Cham, Switzerland: Springer, 2018, pp. 234–244.

[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[26] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[27] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.

[28] Y. Le Cun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel, "Handwritten digit recognition with a back-propagation network," in *Proc. 2nd Int. Conf. Neural Inf. Process. Syst.*, 1989, pp. 396–404.

[29] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep Learning*, vol. 1. Cambridge, MA, USA: MIT Press, 2016.

[30] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, no. 6583, pp. 607–609, Jul. 1996.

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017, *arXiv:1706.03762*.

[32] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," 2021, *arXiv:2101.01169*.

[33] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[34] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," 2020, *arXiv:2012.12877*.

[35] J. Lei Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.

[36] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Cham, Switzerland: Springer, 2018, pp. 3–11.

[37] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.

[38] O. Oktay, J. Schlemper, L. Le Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention U-Net: Learning where to look for the pancreas," 2018, *arXiv:1804.03999*.

[39] A. G. Roy, N. Navab, and C. Wachinger, "Recalibrating fully convolutional networks with spatial and channel 'squeeze and excitation' blocks," *IEEE Trans. Med. Imag.*, vol. 38, no. 2, pp. 540–549, Aug. 2018.

[40] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[41] H. Dou, D. Karimi, C. K. Rollins, C. M. Ortinau, L. Vasung, C. Velasco-Annis, A. Ouaalam, X. Yang, D. Ni, and A. Gholipour, "A deep attentive convolutional neural network for automatic cortical plate segmentation in fetal MRI," *IEEE Trans. Med. Imag.*, vol. 40, no. 4, pp. 1123–1133, Apr. 2021.

[42] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *Proc. Artif. Intell. Statist.*, 2015, pp. 562–570.

[43] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 565–571.

[44] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, 2014, pp. 1–13.

[45] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[46] D. Karimi, S. K. Warfield, and A. Gholipour, "Transfer learning in medical image segmentation: New insights from analysis of the dynamics of model parameters and learned representations," *Artif. Intell. Med.*, vol. 116, Jun. 2021, Art. no. 102078. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0933365721000713

[47] W. Bai, O. Oktay, M. Sinclair, H. Suzuki, M. Rajchl, G. Tarroni, B. Glocker, A. King, P. M. Matthews, and D. Rueckert, "Semi-supervised learning for network-based cardiac mr image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2017, pp. 253–260.

[48] P. Bilic *et al.*, "The liver tumor segmentation benchmark (LiTS)," 2019, *arXiv:1901.04056*.

[49] N. Heller, N. Sathianathen, A. Kalapara, E. Walczak, K. Moore, H. Kaluzniak, J. Rosenberg, P. Blake, Z. Rengel, M. Oestreich, J. Dean, M. Tradewell, A. Shah, R. Tejpaul, Z. Edgerton, M. Peterson, S. Raza, S. Regmi, N. Papanikolopoulos, and C. Weight, "The KiTS19 challenge data: 300 kidney tumor cases with clinical context, CT semantic segmentations, and surgical outcomes," 2019, *arXiv:1904.00445*.

[50] I. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit, M. Lucic, and A. Dosovitskiy, "MLP-mixer: An all-MLP architecture for vision," 2021, *arXiv:2105.01601*.

[51] E. Voita, D. Talbot, F. Moiseev, R. Sennrich, and I. Titov, "Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned," 2019, *arXiv:1905.09418*.

[52] M. Behnke and K. Heafield, "Losing heads in the lottery: Pruning transformer attention in neural machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 2664–2674.

**DAVOOD KARIMI** received the Ph.D. degree in electrical and computer engineering from The University of British Columbia (UBC), Canada. His Ph.D. dissertation was focused on image reconstruction and enhancement for cone-beam computed tomography. After completing his Ph.D. degree, he worked as a Postdoctoral Research Fellow at UCLA and UBC mostly on projects centered on developing machine learning-based methods for medical image segmentation and cancer detection and grading in digital histopathology. He is currently an Instructor in radiology with the Harvard Medical School, and a Scientist with the Computational Radiology Laboratory, Radiology Department, Boston Children's Hospital. His research at IMAGINE involves development of new deep learning algorithms for medical image analysis, including improved algorithms and techniques for motion-robust fetal and newborn imaging for the analysis of early brain development.

**HAORAN DOU** received the B.Eng. degree from Sichuan University, in 2017, and the M.Eng. degree from Shenzhen University. He is currently pursuing the Ph.D. degree with the Center for Computational Imaging & Simulation Technologies in Biomedicine, School of Computing, University of Leeds. His current research interests include medical image segmentation and generative model.

**ALI GHOLIPOUR** (Senior Member, IEEE) received the B.S. and M.S. degrees in electrical engineering from the University of Tehran, in 2001 and 2003, respectively, and the Ph.D. degree in electrical engineering from the University of Texas at Dallas, in 2008. He is currently an Associate Professor in radiology with the Harvard Medical School and the Director of the Intelligent Medical Imaging Research Group and the Translational Research with the Radiology Department, Boston Children's Hospital. His research interests include medical imaging and image analysis, in particular on the development of new imaging technologies and methods to study early human brain development.