An Exploration of Transformer and Convolution Layers in Medical Image Segmentation

by

Xiyao Fu

Bachelor, University of Electronic Science and Technology of China, 2017

Submitted to the Graduate Faculty of

the Swanson School of Engineering in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2023

UNIVERSITY OF PITTSBURGH

SWANSON SCHOOL OF ENGINEERING

This thesis was presented

by

Xiyao Fu

It was defended on

April 27, 2023

and approved by

Zhi-Hong Mao, Ph.D., Professor, Department of Electrical and Computer Engineering

Ahmed Dallal, Ph.D., Assistant Professor, Department of Electrical and Computer Engineering

Thesis Advisor: Liang Zhan, Ph.D., Associate Professor, Department of Electrical and Computer Engineering

An Exploration of Transformer and Convolution Layers in Medical Image Segmentation

Xiyao Fu, M.S.

University of Pittsburgh, 2023

Deep convolutional neural networks (DCNNs) are a popular deep learning technique that has been widely used in segmentation tasks and has received positive feedback. However, DCNN-based frameworks are known to be inadequate in dealing with global relations within imaging features when it comes to segmentation tasks. While several techniques have been proposed to enhance the global reasoning of DCNN, these models are either unable to achieve satisfactory performance compared to traditional fully-convolutional structures or unable to utilize the fundamental advantages of CNN-based networks, namely the ability of local reasoning. In this study, we designed a novel attention mechanism for 3D computation and used it to fully extract the self-attention ability. We proposed a new segmentation framework (called 3DTU) for three-dimensional medical image segmentation tasks, which processes images in an end-to-end manner and performs 3D computation on both the encoder side (with a 3D transformer) and the decoder side (based on a 3D DCNN). In comparison to existing attempts to combine FCNs and global reasoning methods, our framework outperforms several state-of-the-art segmentation methods on two independent datasets consisting of 3D MRI and CT images, as evidenced by experimental results.

Table of Contents

List of Tables

List of Figures

Preface

I would like to express my heartfelt gratitude to all those who have helped me throughout my Master's program and the completion of this thesis. First and foremost, I am indebted to my thesis advisor, Dr. Liang Zhan, for his unwavering support, invaluable guidance, and mentorship throughout the entire process. Dr. Zhan's expertise, dedication, and insights were instrumental in the development of my thesis and in defining the research problem. I am genuinely grateful for his patience, encouragement, and constructive criticism which helped me to improve my work continuously.

I would also like to extend my sincere thanks to Dr. Haoteng Tang for his invaluable contributions to this research project. Dr. Tang's help in writing the paper and designing the experiments was crucial in the successful completion of this thesis. His daily input, support, and critical feedback were instrumental in refining my ideas and methodologies. I am truly grateful for his unwavering support and dedication to this research project.

From my preliminary exam in 2022 until this thesis, I published several papers under the guidance of Dr. Zhan and Dr. Tang, which all made considerable contributions to medical image segmentation.

Lastly, I would like to express my gratitude to all the faculty members and staff who have contributed to my education and research during my time in the Master's program. Their guidance, support, and expertise have been invaluable in shaping my research skills and preparing me for the next chapter of my academic journey.

Thank you all for your valuable contributions and for making my Master's program a fulfilling and enriching experience.

1.0 Introduction

In the past few years, deep convolutional neural networks (DCNNs)[23, 38, 16, 33, 1, 18] have achieved considerable progress in medical image segmentation [4, 30, 31, 32, 40, 48]. However, limited to the local receptive field of the convolutional filter, DCNN-based frameworks are incapable of capturing long-range dependencies from global features for semantic segmentation. To tackle this, several strategies can be considered. First is to use the dilated convolution operation to enlarge the size of the receptive field of the convolutional filter [46, 49, 45, 29]. However, this enlarged local receptive field is still limited by the size of dilation. Another solution is to model the feature map as graph structures and investigate the long-range dependencies through the message-passing mechanism of different graph learning models (e.g., graph convolution networks) [26, 6, 25, 21]. Although these graph learning models have shown great potential in enhancing the global reasoning ability of DCNNs, they have very high requirements for computation and memory due to the constructed large-size graphs.

The attention mechanism [42, 17] is a computation scheme that tries to generate representations via different types of global features at each step. Since attention can be regarded as the conversion and transformation among the query(q), key (k), and value (v) triplet, attention computation is to generate the q based on the combination of the k-v pair. As it is natural to integrate a cycling computation in recurrent cells, traditional attention mechanisms are integrated within recurrent neural networks (e.g.,[17, 7]), which inevitably impairs the efficiency of recurrent networks compared with linear/residual networks [42]. To cope with this, [42] proposed Transformer, a structure consisting of a series of identical encoder blocks connected with a series of identical decoder blocks, which all have no convolutional layer and are connected in a residual way. The original Transformer supported by self-attention works exceptionally well on some tasks like machine translation but not in visual tasks [3]. This is mainly due to the lack of convolution layers which makes the model struggle to detect local features.

For the aforementioned reasons, convolutional-based frameworks are still preferred for

segmentation tasks. Although several other models [5, 13] have been proven feasible, DCNNs remain to be one of the most effective methods. Multiple variants of DCNNs have been proposed to make the segmentation process more effective, one of the most crucial ones is the UNet [37], which is a symmetric structure consisting of convolutional blocks with skip-connections. These convolutional blocks have descending dimensions on the encoder side and ascending dimensions on the decoder side. However, due to the intrinsic fully-convolution structure, UNet is suboptimal to relate local features to global representations with more variant distribution [3]. To cope with the drawbacks of UNet, many methods have been proposed [28, 51, 18, 10]. However, these methods are either very time-consuming or require heavy computations which makes it impossible to be applied to 3D objects.

Under such circumstances, the self-attention mechanism seems to be a nearly optimal solution. It is highly modulized and can stretch the number of self-attention cells according to the training environment. It can also train on vast datasets since the training nature of attention. Therefore, researchers combined the Transformer with convolutional layers for medical image segmentation [24]. On one hand, the Transformer encodes tokenized image patches from a CNN feature map as the input sequence for extracting global contexts. On the other hand, the decoder upsamples the encoded features which are then combined with high-resolution CNN feature maps to enable precise localization.

However, this approach still has some obstacles, especially in the segmentation of 3D objects. This is partially due to Transformers[42] require the input features to have temporal information. Since self-attention does not compute with a clear direction, features have to be preprocessed with temporal info (e.g., cosine function) as input embeddings before training. Although this learning process can be seen as natural (scanning the features linearly and with order), it will restrict the performance of high-dimensional data. For example, many existing Transformer approaches [3, 34, 18] will cut the 3D object into 2D slice sequences to meet the temporal encoding requirement, the segmentation performance, however, is actually worse, which may be due to the 2D slice cutting will destroy the smoothness of the object in 3D space. Bi-directional Transformer [9] is a powerful upgrade version of Transformer. It is a structure with no decoder and processes the inputs all at once with masks to create temporal/spatial continuity. However, we will show in the experiment section that bi-directional

Transformers can serve as a strong encoder but still struggles to get better results on 3D segmentation. To compensate for the loss of feature resolution brought by Transformers, we propose 3D Transformer UNet (3DTU), which employs a hybrid CNN-Transformer architecture to leverage both detailed high-resolution spatial information from CNN features and the global context encoded by our new 3D bi-directional Transformer module. We show that such a design allows our framework to preserve the advantages of self-attention mechanisms and also get considerably improved results on 3D image segmentation compared with previous U-Net-based or Transformer-based methods. To sum up, our contributions in this paper can be summarized as follows:

- We proposed a new 3D bi-directional framework to learn deep 3D features for medical image semantic segmentation.
- We designed a novel attention mechanism specifically suitable for network training and self-attention computation for 3D objects.
- We verified our new framework on multiple datasets, consisting of different imaging modalities (MRI and CT images) and different organs (placenta and lungs infected with COVID) and got state-of-the-art (SOTA) results. Our method beat baselines in performances on multiple metrics.

2.0 Related Work

2.1 Fully Convolutional Network in Medical Image Segmentation

Many studies have attempted to adopt convolutional networks to medical image segmentation. For example, [28] presented a hybrid network consisting of both 3D CNN and 2D CNN in brain image segmentation for Alzheimer's Disease (AD) studies. [37] presented Unet, one of the most iconic encoder-decoder-based methods for medical image segmentation. Their method consists of convolutional blocks that have a U-shaped dimension variation. Specifically, from the input layer of the encoder to the input layer of the decoder, each block's dimension is descending. And the decoder has an ascending dimension that is matched to the encoder blocks. Such a design makes sure that the learning ability of the framework is powerful enough to find the abstract of the locality and output a global representation map. Several adjustments (e.g., [51, 18]) have been made to the original UNet model. For example, U-Net3+ [18] and its variations, although proved effective, still suffer from the locality-heavy learning scheme. Some researchers tried to boost the local reasoning of convolutional layers through the residual structure. For example, ResUNet[10] proposed a residual block between every two convolutional blocks in both the encoder side and decoder side as well as skip-connection between residual blocks with the same dimension between the encoder and decoder. [19] argued that the understanding of the datasets needed in training is more important than the network itself since most UNet-based moderations have achieved little progress. They proposed nnUNet, a robust network that is designed based on the combination of 2D and 3D UNet. They also made different training configurations (normalization tricks, cropping, activation functions, etc.) based on the datasets.

2.2 Transformers

Transformers [42] were initially proposed for general NLP tasks and quickly gain widespread attention by beating previous most state-of-the-art results by a large margin. [9] converted the original Transformer model into BERT introduced so-called bi-directional Transformers and is proven effective again. Naturally, multiple efforts have been made to adjust the learning ability of Transformers in the computer vision domain. Several variants of Transformers have emerged recently. [34] is one of the early works to adjust vanilla Transformers by incorporating visual information. This model pre-processes each pixel of one image through a 1×1 convolution layer. Then the embeddings are computed with positional embeddings before feeding into Transformers for super-resolution tasks. In another attempt for visual tasks, [11] proposed Vision Transformer (ViT), which presented a novel way of input embedding on visual information. It achieved state-of-the-art on ImageNet classification by directly applying Transformers with global self-attention to full-sized images. Specifically, ViT flattens an image to fixed-sized pixels which then be linearly added to positional embeddings before feeding to Transformer encoders. [41] presented gated axial attention that creates a gated scheme to improve learning ability on the local scale.

2.3 Combination of UNet and Transformer in Medical Image Segmentation

Multiple attempts have been made to combine the UNet with Transformer in both framework structure and inner encoder/decoder computation. TransUNet [3] consists of a series of Transformer units as the encoder and the right half of the UNet as the decoder to generate predictions in medical image segmentation. Both the encoder and the decoder in [3] are computed in a 2D scenario. [47] introduced SpecTr, a framework that takes spectral normalization into the computation between convolution and attention blocks. Their methods achieved better results than the baseline when training on hyperspectral medical images. [43] presented TransBTS which utilizes 3D CNN to extract input representations. UNet Transformer, presented by [35], replaces self-attention modules in Transformer encoder/decoder

cells by convolutional blocks and batch normalization computations. Another attempt is Swin-UNet [2], which instead replaces convolution blocks in the UNet-Structure network with self-attention modules. Several works follow the similar manners including UNETR [15], SWIN UNETR [14], CoTr [44], nnFormer [50], DS-TransUNet [27], UTNet [12] and PNS-Net [20], etc. In UNETR, the authors presented a novel 3D Transformer encoder and a voxel-wise loss for model training. For the positional embedding, they adopted a strategy from the Visual Transformer which divides the 3D images into 3D patches. The decoder in their work consists of several convolutional blocks in different dimensions and skip connections to the encoder. The SWIN UNETR is proposed for $3D$ multi-modal MRI brain image studies, which is different from the SWIN UNET that is proposed for 2D images. The CoTr utilized a DeTrans-encoder with a novel attention mechanism and a CNN-based decoder. The nnFormer utilizes CNN as part of the encoder, which leverages the ability of local feature extraction of CNN structures. Moreover, it utilizes transformer structures as its decoder and the second part of its encoder. There are two differences between our 3DTU and the nnFormer. First, we utilize a CNN-based structure (i.e., the right part of 3DUNet) as our decoder. And we design an attention mechanism that computes the attention scores from different directions.

The aforementioned methods adjust the Transformers in visual tasks by introducing their own positional embedding rules. Although these rules are to an extent useful, their performance all suffers from the slicing of 3D data to adjust the positional embeddings. In our paper, positional embeddings are not needed technically, even for 3D data. We modify the multi-head attention from its original form to a refined computation scheme that fully utilizes the potentials of Transformer and UNet. More importantly, our encoder is a refined bi-directional Transformer, which learns the feature from three (i.e., along x, y, and z) directions simultaneously [1].

[1]We use the term 'bi-directional' by following previous studies. However, our 3DTU learns the features from three directions instead.

3.0 Methods

We propose a 3D Unet-based framework with bi-directional transformers (named 3DTU) in this work. The self-attention mechanism in the proposed bi-directional transformers can improve the ability to generalization of the framework encoder. We will delve into the technical details in this section.

As shown in Figure 1, our proposed 3DTU is an encoder-decoder framework, where the encoder consists of two modules including a feature extraction module (see Part I in Figure 1) and a bi-directional transformer module (see Part II in Figure 1). Given a 3D image $I \in \mathcal{R}^{h \times w \times d \times c}$, where h, w and d is the shape of the image and c is the image channel number, the feature extraction module projects the 3D image I as a latent representation X via basic convolutional neural networks (CNNs). Then the 3D bi-directional transformer cells take the latent representation X as input and yield the masked latent representation X_M by using Masked-LM (MLM) [9] step by step. Finally, the decoder part utilizes the masked latent representations to reconstruct the segmentation predictions for loss computation.

3.1 Encoder with 3D Bi-directional Transformer

As aforementioned, the encoder of the 3DTU consists of two parts. The first part of the encoder is a CNN-based feature extraction module. We aim to convert the original 3D image (I) into an iso-dimensional latent cube representation ($X \in \mathcal{R}^{1 \times p \times p \times p}$) via this module as assistance to capture the image locality for transformer modules, since the transformer module may not have enough ability to capture the image local features. We will show this point in the ablation studies. Particularly, the feature extraction module includes two convolutional layers followed by a fully-connected (FC) layer, and a max-pooling layer in between the two convolutional layers. The FC layer is used to adapt the feature dimension.

The bi-directional transformer module takes the latent cube representation X as input and computes multi-head attentions with the MLM strategy [9]. Details of the bi-directional

transformer module are shown in Figure 2. In general, each cell in the bi-directional transformer module generates the latent feature map X_1 by the following steps:

$$X^{'} = Att(Norm(X)) + X, \tag{3-1}$$

$$X^{''} = FF(Norm(X^{'})), \tag{3-2}$$

$$X_1 = X^{'} + X^{''}, \tag{3-3}$$

where $Att(\cdot)$ is the multi-head self-attention operation, $Norm(\cdot)$ is a 3D normalization operation, and $FF(\cdot)$ is the feed forward layer (i.e., FC layer). + denotes a pixel-wise add operation. Particularly, the multi-head attention is computed by:

$$Att\_head_i^{x,y,z} = SDP(Q, K, V) \times W, \tag{3-4}$$

$$MultiHead(Q, K, V) = Concat(head_i^x, head_i^y, head_i^z), \tag{3-5}$$

where $SDP(\cdot)$ is the Scaled Dot-Product Attention, W is the trainable parameters for linear projections (i.e., L_q, L_k, L_v in Figure 2) and $Concat(\cdot)$ denotes a concatenation operation. Q, K and V are the query-key-value triplets defined by the transformer cell. Note that our proposed attention mechanism can yield the attention score by scanning the query-key-value triplets in 3 different directions (i.e., along x, y, and z axis, respectively), which gains plentiful discriminative and anisotropic semantic information for 3D image segmentation.

3.2 UNet-based Decoder

As shown in Figure 1, we utilize convolutional blocks with ascensional dimensions in the decoder part. A residual connection is adopted between the encoder side and the decoder side. Particularly, a cascaded of multi-channel feature map (FM) blocks are integrated into the decoder part, each of which contains two $3 \times 3 \times 3$ convolutional layers and an upsampling layer. The channel number of feature maps reduces by half after each FM block. In the last FM block, instead of upsampling layer, a $1 \times 1 \times 1$ convolutional layer is used to generate final segmentation predictions.

3.3 Loss Function and Supervision Manner

Since the MLM strategy is used in the encoder part, where a portion of image features are masked (i.e., set to 0 values) and the other portions remain the same. Hence, our goal is to use the uncovered portions to predict the masked portions [9], which results in that the loss is only estimated based on the masked regions. Particularly, the loss function can be formulated as:

$$\mathcal{L} = \alpha \times \ell_{dice}(\hat{y}_{mask}, y_{mask}) + (1 - \alpha) \times \ell_{BCE}(\hat{y}_{mask}, y_{mask}), \tag{3-6}$$

where $\hat{y}_{mask}$ and y_{mask} are the masked regions of segmentation prediction and ground truth, respectively. $\alpha \in [0, 1]$ is the loss weight.

4.0 Experiments

4.1 Datasets

We used three datasets obtained from different modalities for this study, including Placenta MRI (Placenta) dataset, and COVID-19 CT lung and infection segmentation (Covid20) dataset, as well as Multi-Atlas Labeling Beyond the Cranial Vault (Synapse) dataset. Details of data description and preprocessing are shown below.

Placenta MRI Dataset was collected from the Washington University in Saint Louis (WUSTL) [39], where all data were de-identified before processing. The data collection and related studies were approved by the Institutional Review Board at the WUSTL. 81 MRI scans were collected from 46 pregnant patients (mean age = 23.91 ± 3.02 yo, mean BMI = 25 ± 3.66 at recruitment) with normal singleton pregnancy underwent MRI during the third trimester, by a Siemens 3T VIDA scanner. 21 of 46 patients had the single scan and 25 patients had multiple longitudinal scans. The average gestational ages (GA) during MRI scans were 34.12 ± 1.07 weeks (Min GA 28 wk 3 days, max GA 38 wk 6 days). T2-weighted MRI of the entire uterus was acquired with a 2D EPI sequence in the left lateral position. The MRI data has a fixed acquisition matrix of $128 \times 128 \times 115$, and variable voxel sizes from $3 \times 3 \times 3$ mm to $3.5 \times 3.5 \times 3.5$ mm, up to the patients' size. Manual segmentation of the placenta regions was conducted by experienced radiologists for all MRI images.

COVID19-CT-Seg20 Dataset (Covid20) contains 20 COVID-19 3D CT images, where lungs and infections were annotated by two radiologists and verified by an experienced radiologist [1] [22]. We only focused on the segmentation of the COVID-19 infections in this study, since it is more challenging and important.

Multi-Atlas Labeling Beyond the Cranial Vault (Synapse) Dataset. [2] We use the 30 abdominal CT scans from the MICCAI 2015 Multi-Atlas Abdomen Labeling Challenge. These scans were captured during portal venous contrast phase with variable

[1] `https://zenodo.org/record/3757476#.Y1NGmy1h1B1`

[2] `https://www.synapse.org#!Synapse:syn3193805/wiki/217789`

volume sizes (512 x 512 x 85 - 512 x 512 x 198) and field of views (approx. 280 x 280 x 280 mm^3 - 500 x 500 x 650 mm^3). The in-plane resolution varies from 0.54 x 0.54 mm^2 to 0.98 x 0.98 mm^2, while the slice thickness ranges from 2.5 mm to 5.0 mm. we report the average experimental results on 8 abdominal organs (aorta, gallbladder, spleen, left kidney, right kidney, liver, pancreas, spleen, stomach) with five-fold validation.

4.2 Implementation Details

In the pre-processing step, we simply normalized the intensities of each 3D image to zero mean and unit variance. In the training phase, we applied data augmentation techniques to reduce potential overfitting, including random rotation the image by 90° along three dimensions, and adjusting the brightness of the top 3% pixels. The training iterations were set to 10^5. We trained the model using the Adam optimizer with a batch size of 1 and synchronized batch normalization. The initial learning rate was set to $1e-2$ and was decayed by $(1-\frac{current\_epoch}{max\_epoch})^{0.9}$. We also regularized the training with dropout in the transformer cells. All experiments are conducted using a five-fold cross-validation, based on Pytorch 1.7.1 on a workstation with 2 NVIDIA TITAN RTX GPUs. The data division on the public Covid20 dataset is adopted by following the division strategy in [36].

As aforementioned, our encoder consists of two parts. In the feature extraction module, we used a CNN network with two conv layers, one max-pooling layer, and one $1-$D fully-connected layer with the direction of $x-y$ plane to z coordinate to convert the representations with the original dimension to a cube. The first cov layer, with a kernel size of $3\times3\times3$, embeds the input $3-$D image into local representation maps, while the second conv layer project the local representation maps for the second part of the encoder via a linear transformation. The output dimension of the feature extraction module is converted (i.e., reshape) to $X \in \mathcal{R}^{1\times256\times256\times256}$. In the bi-directional transformer module, we utilize multiple transformer cells with the bi-directional self-attention mechanism. Specifically, the input embedding strategy that we adopted is Masked LM (MLM) [9]. The Masked LM has been proved to be useful within the previous BERT paper [42], where the image portion masked in the

encoder is matched to that in the loss computation stage. Moreover, since we do not embed the data with the positional encoding in our framework, we require a way to learn the 3D representations through a certain sequence. MLM can well meet this requirement. We set the number of transformer cells as 12, 6, and 6 for Placenta, Covid20, and Synapse datasets, respectively. The number of heads within each transformer cell is 15, where each direction (i.e., $x-y$, $x-z$ and $y-z$ plane) contains 5 heads to compute self-attention scores. The length of each mask is set to 16, 32 and 32 for the Placenta, Covid20, and Synapse dataset, respectively. Each cube representation is divided into 16 parts in the training phase.

4.3 Baseline Settings and Evaluation Metrics

To evaluate our 3DTU's performance, we choose the following frameworks as baselines: 2DU-Net[37], 3D U-Net[8], UNet++[51], TransUNet[3] as well as ViT (visual transformer)[11], nnFormer[50], nnUNet[19]. Both 2D and 3D UNet are FCN-based encoder-decoder structures with convolutional blocks and skip-connections between the encoder and decoder. The UNet++ is a nested-connected encoder-decoder structure, where each convolutional block is connected to all other blocks. The TransUNet is an encoder-decoder network, where the encoder of UNet is replaced by a 2D transformer including a positional embedding scheme followed by Visual Transformer (ViT). The nnFormer is a 3D UNet-type framework which replaces the conv blocks by three different novel attention mechanisms.

The metrics we used to evaluate our 3DTU include mIoU, DICE score and Hausdorff Distance (HD). IoU is the area of overlap between the predicted segmentation and the ground truth divided by the area of union between them. For binary (two classes) or multi-class segmentation, the mean IoU (mIoU) of the image is calculated by taking the IoU of each class and averaging them. DICE score is the harmonic mean of precision and recall of the segmentation results. mIOU and DICE scores are two overlap-based metrics measuring the similarity between the ground truths and segmentation predictions. The range of mIOU and DICE scores is from 0 to 1 and the larger value indicates better segmentation performance. The directed average Hausdorff distance (HD) from point set X to Y is computed by the sum

of all minimum distances from all points from point set X to Y divided by the number of points in X. HD is a shape distance-based metric, which measures the dissimilarity between the surfaces of the segmentation results and the related ground truths. A lower value of HD indicates better performance.

4.4 Comparative Experiments

Table 1 provides the performance of our proposed 3DTU and the six competing baselines, including 2D UNet [37], 3D UNet[37], UNet++ [51], TransUNet [3] and visual transformer (ViT) [11], and nnFormer [50] on the Placenta and Covid20 datasets. It shows that our 3DTU outperforms all competing baseline methods consistently in terms of mIOU and DICE scores on both datasets, while beating most of the methods in the baseline in Synapse dataset, indicating that the segmentation results of our models match well with the ground-truth. For example, our proposed 3DTU outperforms baselines with at least 0.48% and 0.44% increases in DICE scores on Placenta and Covid20 datasets, respectively. This may attribute to the attention mechanism proposed in the 3DTU which can compute the attention scores from three different directions to yield discriminative and anisotropic semantic features for 3D images. In general, the transformer-based methods (e.g., TransUNet, ViT etc.) perform better than the other baseline methods. In addition, we visualized the segmentation results of our 3DTU and the best baseline method (i.e., nnUNet) on three datasets in Figure 4, Figure 5 and Figure 6, respectively.

4.5 Ablation Study

We conducted an ablation study on both datasets (i.e., Placenta and Covid20) to evaluate the effectiveness of each part in our 3DTU framework. Our 3DTU is an encoder-decoder-based framework, where the encoder consists of a CNN-networks part as well as a bi-directional transformer (BiT) part, where the decoder is in the UNet decoder setting.

Hence, we designed the following four experiments in our ablation study.

- We removed the CNN networks in the encoder and directly fed the input images to the BiT part.
- We removed the BiT part in the encoder and directly connected the CNN networks to the UNet decoder.
- We removed the UNet decoder part and consider the BiT as both (part of) encoder and decoder [3].
- We designed a comparative experiment where we train 3DTU with positional encoded representations. We encoded the representations at the input of the Transformer encoder.

The results in Table 2 show the effectiveness and necessity of all the sub-parts in our 3DTU. The results in Table 3 indicate that the positional encoding is not necessary for our framework since our attention mechanism can process the 3D data as a whole. Comparing with the 3DTU w/o positional encoding, the segmentation dice scores yielded by 3DTU with positional encoding are not changed or even decreased. When we removed the CNN-networks and only utilized BiT as the encoder (see results of BiT+Unet Decoder in Table 2), the segmentation performance decrease on both datasets (e.g., DICE decrease from 84.0% to 66.9% and from 92.0% to 72.8% on Placenta and Covid datasets, respectively). This indicates an essential role of CNN-based conv layers in the encoder, without which the self-attention transformer layers may not localize the raw image pixels precisely. Meanwhile, the segmentation performance increase when we use BiT instead of UNet as decoder (see results of CNN + UNet Decoder and CNN + BiT). This manifests that, compared with UNet-based methods, the (bi-directional) transformers are more powerful in boosting the segmentation results.

4.6 Parameter Analysis

We analyze the impact of two parameters, including the loss weights α and the number of transformer cells, on the segmentation performance of our proposed 3DTU across two

[3]It shows in [9] that the bi-directional transformer can serve as both encoder and decoder.

datasets in Figure 3. In general, Figure 3 indicates that the segmentation results performed by our 3DTU are consistent. Figure 3 (A) shows that the dice results increase and then decrease with the increase of α from 0 to 1. The best dice scores are achieved when $\alpha = 0.2$ on both Placenta and Covid20 datasets. Figure 3 (B) shows that the segmentation performance improves when increasing the number of transformer cells from 3 to 6. However, the performance will keep stable (on Placenta dataset) or even slightly decrease (on Covid20 dataset) when the framework goes deeper. The reason of the slight decrease of performance on Covid dataset may result from the small size of dataset. Only 20 3D images are included in Covid20 dataset, which may not facilitate the training process when the network goes deeply. Moreover, our 3DTU has a total of 70M parameters (when training on Covid20 dataset and Synapse dataset), which is more than 2D UNet (7M) and 3D UNet (17M) but beats the other transformer-based or hybrid framework in the baseline (the TransUNet has 80M parameters and nnFormer has 158M parameters).

Table 1: Quantitative segmentation results of different methods on two datasets, where mIOU and DICE are in %. The best results are shown in red and the second best results are shown in blue.

.

	Placenta Dataset			Covid20 Dataset			Synapse Dataset		
	mIOU	DICE	HD95	mIOU	DICE	HD95	mIOU	DICE	HD95
2D UNet	67.6	72.3	12.0	73.6	78.3	112.5	56.3	60.6	45.7
3D UNet	72.5	78.6	10.7	78.1	84.0	97.6	59.4	62.2	42.2
UNet++	74.5	77.1	8.2	80.3	84.6	63.0	67.1	73.7	34.0
TransUNet	73.6	80.0	7.4	83.1	89.2	45.8	70.2	77.5	31.7
ViT	72.9	79.7	8.5	84.2	89.0	70.3	65.3	67.9	36.1
nnFormer	78.3	82.1	10.2	81.0	89.9	66.2	81.8	86.6	10.6
nnUNet	78.9	83.6	8.7	90.3	91.6	59.9	84.2	89.8	16.6
3DTU (Ours)	79.8	84.0	7.2	90.5	92.0	59.4	85.0	87.3	18.4

Table 2: Dice scores (in %) of our 3DTU on three datasets. The best results are shown in **bold**.

DICE Score	Placenta Dataset	Covid20 Dataset	Synapse Dataset
CNN + UNet Decoder	68.6	74.3	59.5
BiT + UNet Decoder	66.9	72.8	70.2
CNN + BiT	80.0	89.2	65.1
3DTU	**84.0**	**92.0**	**87.3**

Table 3: Dice scores (in %) of our 3DTU running on data that has been preprocessed with/without positional encoding.

	Placenta Dataset	Covid20 Dataset	Synapse Dataset
3DTU w/o Positional Encoding	84.0	92.0	87.3
3DTU with Positional Encoding	82.7	92.1	86.8

5.0 Conclusion

In this paper, we propose a novel 3D Transformer UNet (3DTU) framework to capture global contextual information for 3D medical image segmentation. A new attention mechanism is proposed with our 3DTU framework, which is especially suitable for computing self-attentions for 3D objects. The experimental results on two 3D medical image datasets demonstrate that our method can outperform several state-of-the-art segmentation baselines. In the future, we plan to explore how to reduce the computation loads in transformer layers, which may improve the efficiency of most current transformer-based methods.

6.0 Data Availability Statement

The Covid20 dataset is from the community of Coronavirus Disease Research - COVID-19 [22] and is available from `https://zenodo.org/record/3757476#.Y1NGmy1h1B1`. The Synapse dataset is available from `https://www.synapse.org#!Synapse:syn3193805/wiki/217789`. The Placenta dataset is available upon request.

7.0 Funding

This project was partially supported by NSF IIS 2045848 and NIH/NICHD (R01HD094381 and R01HD104822), as well as by Burroughs Wellcome Fund Preterm Birth Initiative (NGP10119), and the Bill & Melinda Gates Foundation (INV-005417, INV-035476, and INV-037302).

8.0 Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

9.0 Acknowledgement

We thank the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation (NSF) grant number ACI-1548562 and NSF award number ACI-1445606, which provide the computation resources based on Pittsburgh Supercomputing Center (PSC) for part of our work.

We would like to appreciate the efforts devoted by the community of Coronavirus Disease Research - COVID-19 and Zenodo to collect and share the COVID-19 CT image dataset. Meanwhile, we appreciate the Washington University in Saint Louis to collect and share the data Placenta MRI dataset for our segmentation algorithm evaluations.

Appendix

A.1 Figures

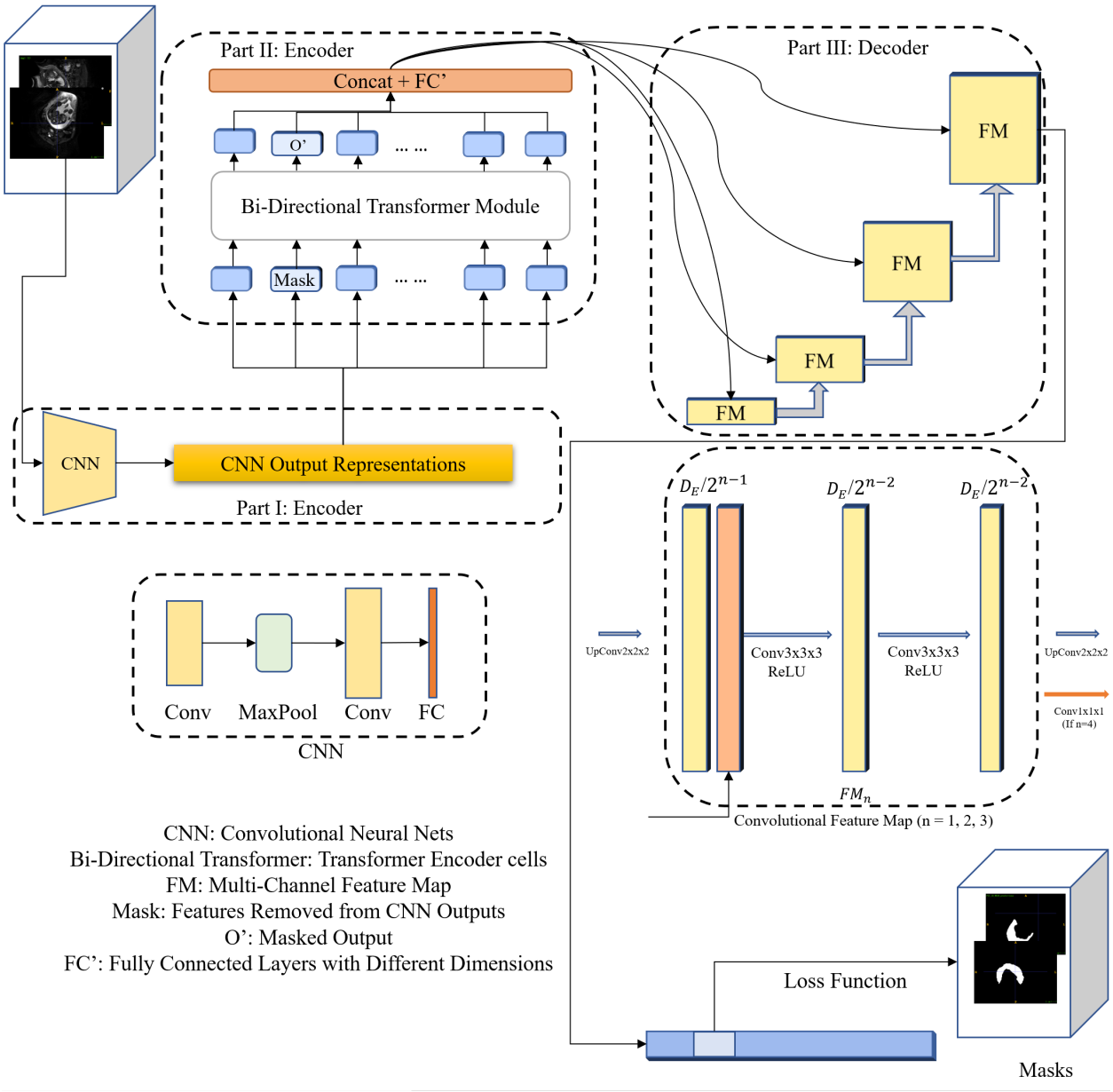

Figure 1: The diagram of the 3DTU framework in an encoder-decoder setting. The encoder consists of two parts including feature extraction and a bi-directional transformer.

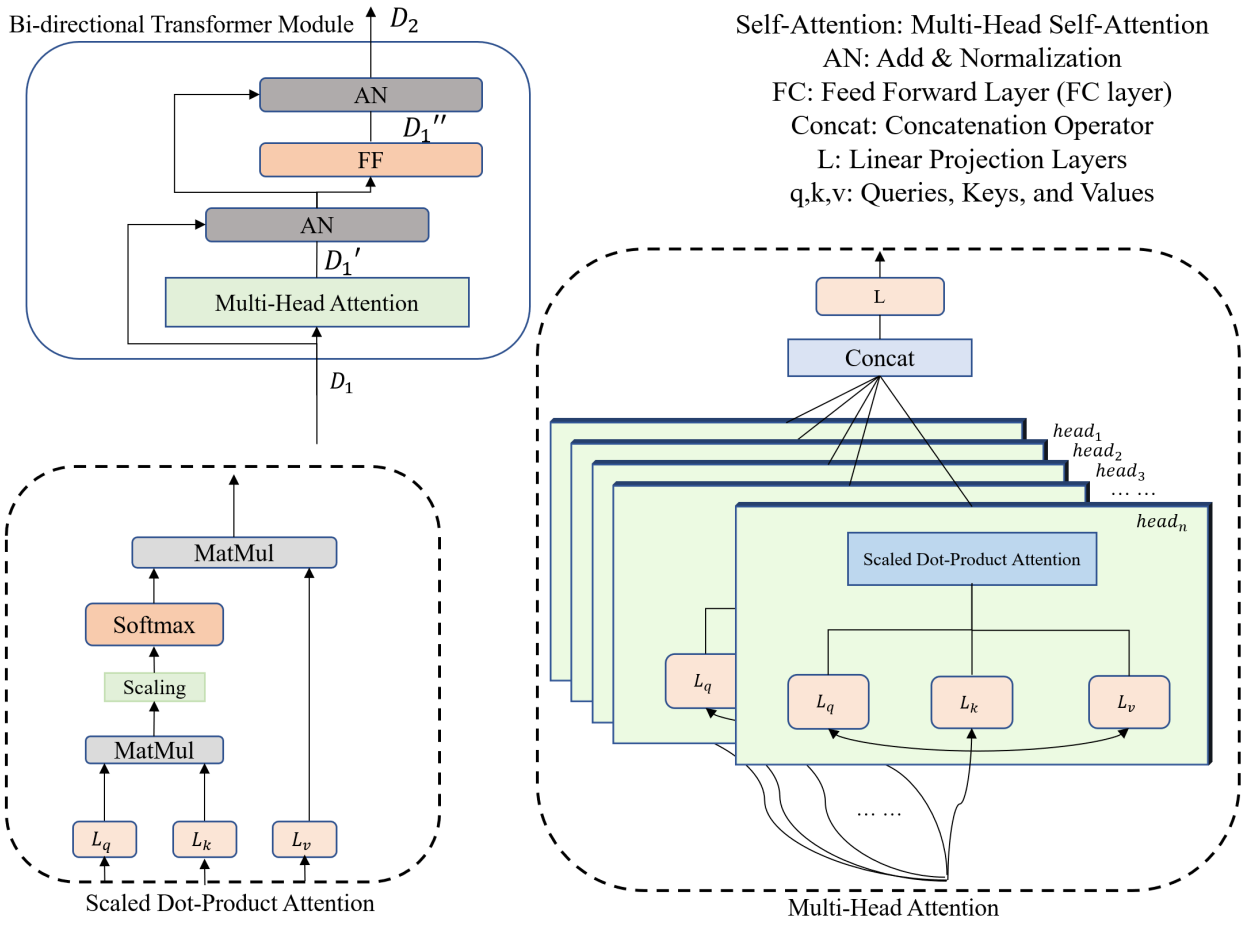

Figure 2: Encoder Part II: bi-directional transformer with a multi-head attention mechanism.

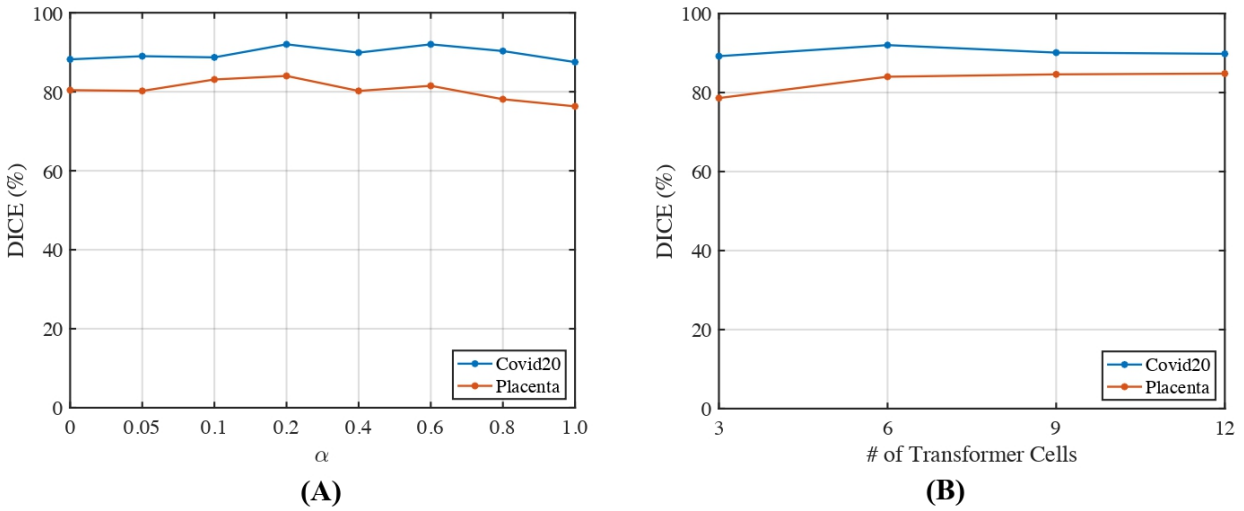

Figure 3: Impacts of α and number of transformer cells on segmentation performance. (A). Dice of 3DTU v.s. α. (B). Dice of 3DTU v.s. the number of transformer cells.

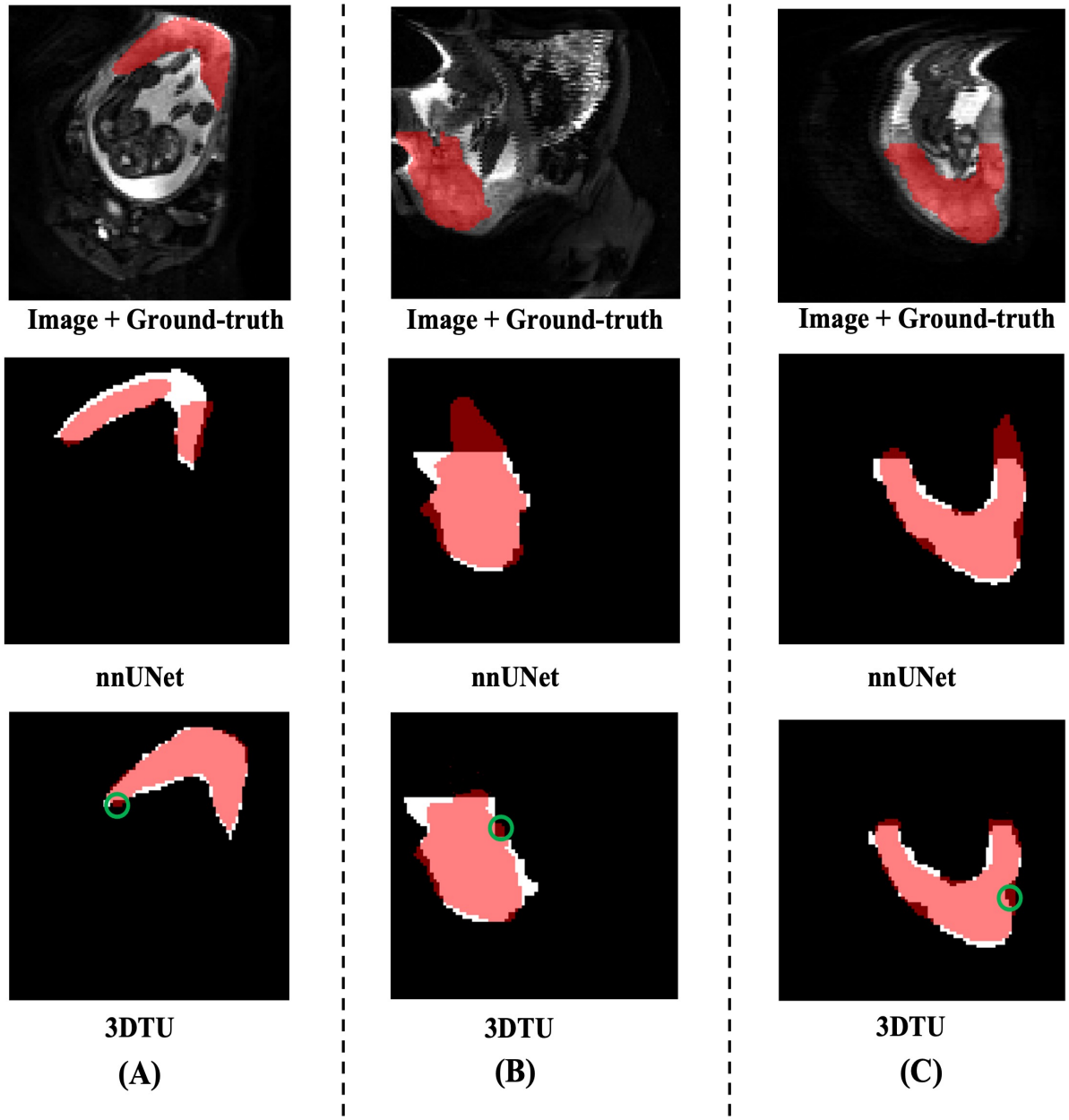

Figure 4: Visualization of the segmentation results on the Placenta dataset produced by our 3DTU and nnUNet. Column (A), (B) and (C) show the x-y plane, y-z plane, and x-z plane of 3D segmentation predictions, respectively. The true-positive regions are highlighted in pink. The false-negative regions are highlighted in red (e.g., the green circle regions in the last row). Better view with colors and zooming in.

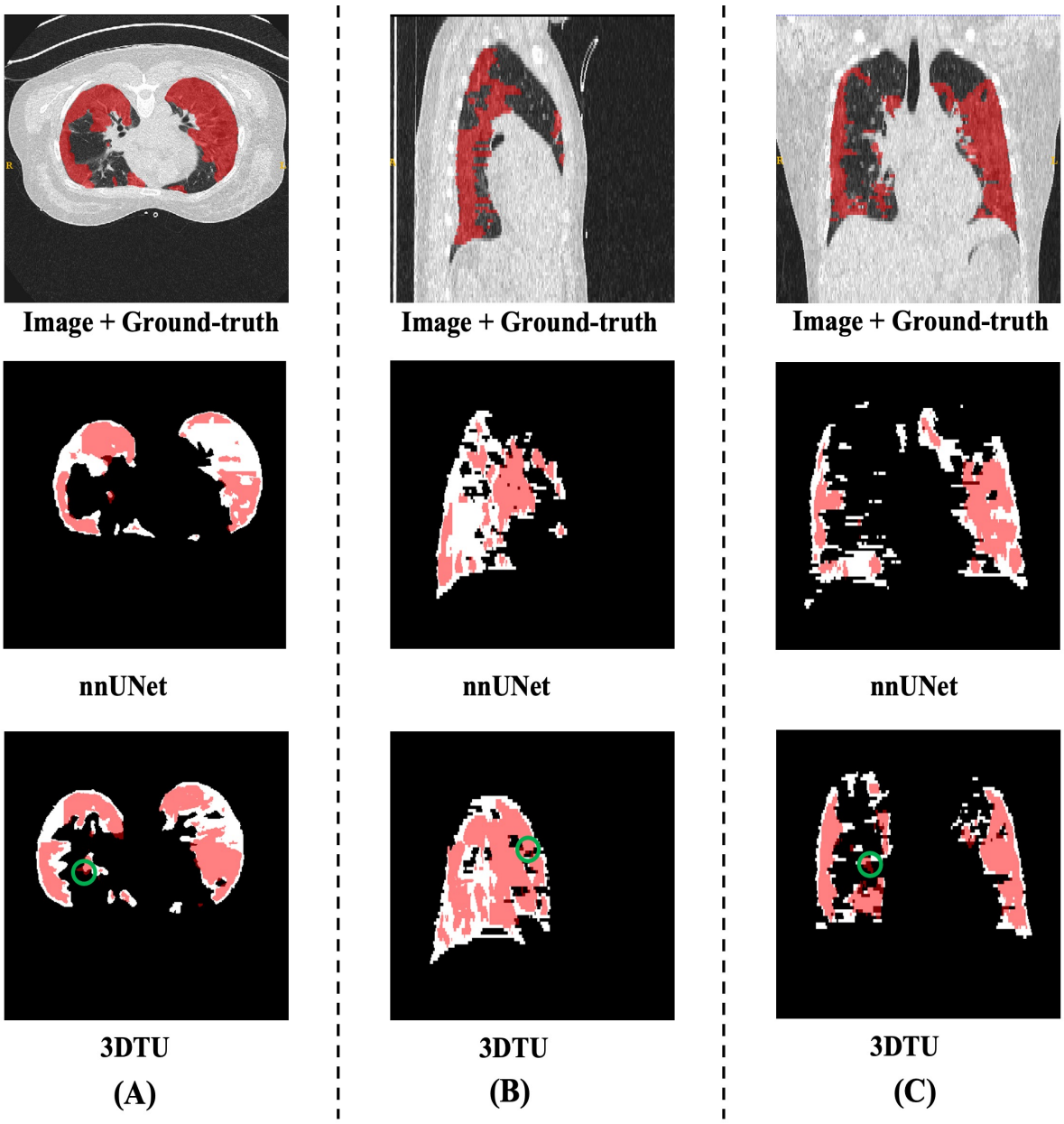

Figure 5: Visualization of the infection segmentation results on the Covid20 dataset produced by our 3DTU and nnUNet. Columns (A), (B), and (C) show the x-y plane, y-z plane and x-z plane of 3D segmentation predictions, respectively. The true-positive regions are highlighted in pink. The false-negative regions are highlighted in red (e.g., the green circle regions in the last row). Better view with colors and zooming in.

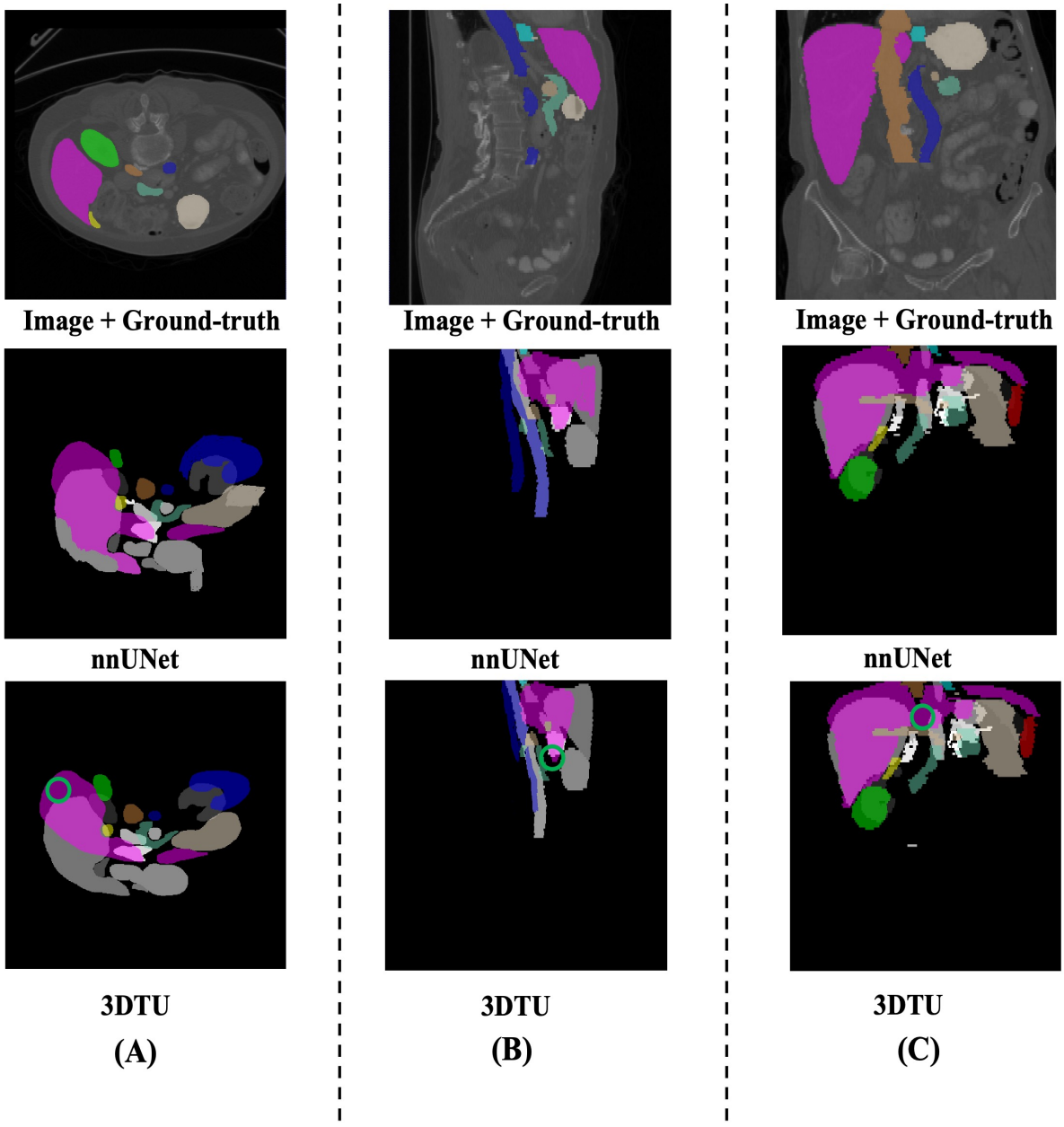

Figure 6: Visualization of the segmentation results on the Synapse dataset produced by our 3DTU and nnUNet. Columns (A), (B), and (C) show the x-y plane, y-z plane, and x-z plane of 3D segmentation predictions, respectively. The green circle indicates part of the false-negative regions. It better view with colors and zooming in.

Bibliography

[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[2] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. *arXiv preprint arXiv:2105.05537*, 2021.

[3] 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.

[4] 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.

[5] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7103–7112, 2018.

[6] Yunpeng Chen, Marcus Rohrbach, Zhicheng Yan, Yan Shuicheng, Jiashi Feng, and Yannis Kalantidis. Graph-based global reasoning networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 433–442, 2019.

[7] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.

[8] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016.

[9] 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.

[10] Foivos I Diakogiannis, François Waldner, Peter Caccetta, and Chen Wu. Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 162:94–114, 2020.

[11] 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.

[12] Yunhe Gao, Mu Zhou, and Dimitris N Metaxas. Utnet: a hybrid transformer architecture for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 61–71. Springer, 2021.

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[14] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. *arXiv preprint arXiv:2201.01266*, 2022.

[15] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 574–584, 2022.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[17] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[18] Huimin Huang, Lanfen Lin, Ruofeng Tong, Hongjie Hu, Qiaowei Zhang, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, and Jian Wu. Unet 3+: A full-scale connected unet for medical image segmentation. In *ICASSP 2020-2020 IEEE Inter-*

national Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1055–1059. IEEE, 2020.

[19] 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.

[20] Ge-Peng Ji, Yu-Cheng Chou, Deng-Ping Fan, Geng Chen, Huazhu Fu, Debesh Jha, and Ling Shao. Progressively normalized self-attention network for video polyp segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 142–152. Springer, 2021.

[21] Haozhe Jia, Haoteng Tang, Guixiang Ma, Weidong Cai, Heng Huang, Liang Zhan, and Yong Xia. Psgr: Pixel-wise sparse graph reasoning for covid-19 pneumonia segmentation in ct images. *arXiv preprint arXiv:2108.03809*, 2021.

[22] Ma Jun, Ge Cheng, Wang Yixin, An Xingle, Gao Jiantao, Yu Ziqi, Zhang Minqing, Liu Xin, Deng Xueyuan, Cao Shucheng, Wei Hao, Mei Sen, Yang Xiaoyu, Nie Ziwei, Li Chen, Tian Lu, Zhu Yuntao, Zhu Qiongjie, Dong Guoqiang, and He Jian. COVID-19 CT Lung and Infection Segmentation Dataset, April 2020.

[23] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[24] Jun Li, Junyu Chen, Yucheng Tang, Bennett A Landman, and S Kevin Zhou. Transforming medical imaging with transformers? a comparative review of key properties, current progresses, and future perspectives. *arXiv preprint arXiv:2206.01136*, 2022.

[25] Xia Li, Yibo Yang, Qijie Zhao, Tiancheng Shen, Zhouchen Lin, and Hong Liu. Spatial pyramid based graph reasoning for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8950–8959, 2020.

[26] Yin Li and Abhinav Gupta. Beyond grids: Learning graph representations for visual recognition. *Advances in Neural Information Processing Systems*, 31, 2018.

[27] Ailiang Lin, Bingzhi Chen, Jiayu Xu, Zheng Zhang, Guangming Lu, and David Zhang. Ds-transunet: Dual swin transformer u-net for medical image segmentation. *IEEE Transactions on Instrumentation and Measurement*, 2022.

[28] Manhua Liu, Danni Cheng, Kundong Wang, and Yaping Wang. Multi-modality cascaded convolutional neural networks for alzheimer's disease diagnosis. *Neuroinformatics*, 16(3):295–308, 2018.

[29] Mengting Liu, Piyush Maiti, Sophia Thomopoulos, Alyssa Zhu, Yaqiong Chai, Hosung Kim, and Neda Jahanshad. Style transfer using generative adversarial networks for multi-site mri harmonization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 313–322. Springer, 2021.

[30] Mengting Liu, Alyssa Zhu, Piyush Maiti, Sophia I Thomopoulos, Shruti Gadewar, Yaqiong Chai, Hosung Kim, Neda Jahanshad, Alzheimer's Disease Neuroimaging Initiative, et al. Style transfer generative adversarial networks to harmonize multi-site mri to a single reference image to avoid over-correction. *bioRxiv*, 2022.

[31] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[32] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1520–1528, 2015.

[33] Xiaoying Pan, Yizhe Zhao, Hao Chen, De Wei, Chen Zhao, and Zhi Wei. Fully automated bone age assessment on large-scale hand x-ray dataset. *International journal of biomedical imaging*, 2020, 2020.

[34] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *International Conference on Machine Learning*, pages 4055–4064. PMLR, 2018.

[35] Olivier Petit, Nicolas Thome, Clement Rambour, Loic Themyr, Toby Collins, and Luc Soler. U-net transformer: Self and cross attention for medical image segmentation. In *International Workshop on Machine Learning in Medical Imaging*, pages 267–276. Springer, 2021.

[36] Yu Qiu, Yun Liu, Shijie Li, and Jing Xu. Miniseg: An extremely minimum network for efficient covid-19 segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 4846–4854, 2021.

[37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[38] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[39] Z Sun, W Wu, P Zhao, Q Wang, P Woodard, DM Nelson, A Odibo, A Cahill, and Y Wang. Dual-contrast mri reveals intraplacental oxygenation patterns, detects placental abnormalities and fetal brain oxygenation. *Ultrasound in obstetrics & gynecology: the official journal of the International Society of Ultrasound in Obstetrics and Gynecology*, 2022.

[40] Hiroki Tokunaga, Yuki Teramoto, Akihiko Yoshizawa, and Ryoma Bise. Adaptive weighting multi-field-of-view cnn for semantic segmentation in pathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12597–12606, 2019.

[41] Jeya Maria Jose Valanarasu, Poojan Oza, Ilker Hacihaliloglu, and Vishal M Patel. Medical transformer: Gated axial-attention for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 36–46. Springer, 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. *Advances in neural information processing systems*, 30, 2017.

[43] Wenxuan Wang, Chen Chen, Meng Ding, Hong Yu, Sen Zha, and Jiangyun Li. Transbts: Multimodal brain tumor segmentation using transformer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 109–119. Springer, 2021.

[44] Yutong Xie, Jianpeng Zhang, Chunhua Shen, and Yong Xia. Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 171–180. Springer, 2021.

[45] Zichao Yang, Zhiting Hu, Ruslan Salakhutdinov, and Taylor Berg-Kirkpatrick. Improved variational autoencoders for text modeling using dilated convolutions. In *International conference on machine learning*, pages 3881–3890. PMLR, 2017.

[46] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

[47] Boxiang Yun, Yan Wang, Jieneng Chen, Huiyu Wang, Wei Shen, and Qingli Li. Spectr: Spectral transformer for hyperspectral pathology image segmentation. *arXiv preprint arXiv:2103.03604*, 2021.

[48] Jianjia Zhang, Luping Zhou, Lei Wang, Mengting Liu, and Dinggang Shen. Diffusion kernel attention network for brain disorder classification. *IEEE Transactions on Medical Imaging*, 2022.

[49] Xiaohu Zhang, Yuexian Zou, and Wei Shi. Dilated convolution neural network with leakyrelu for environmental sound classification. In *2017 22nd international conference on digital signal processing (DSP)*, pages 1–5. IEEE, 2017.

[50] Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnformer: Interleaved transformer for volumetric segmentation. *arXiv preprint arXiv:2109.03201*, 2021.

[51] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE transactions on medical imaging*, 39(6):1856–1867, 2019.