Analyzing Components of a Transformer under Different Data Scales in 3D Prostate CT Segmentation

Yicong Tan

Delft University of Technology

TU Delft

Analyzing Components of a Transformer under Different Data Scales in 3D Prostate CT Segmentation

by

Yicong Tan

to obtain the degree of Master of Science in Computer Science
at the Delft University of Technology,
Faculty of Electrical Engineering, Mathematics and Computer Science

Student Number:	5223245	
Project Duration:	December, 2021 - September, 2022	
Thesis Committee:	Dr. Jan van Gemert,	TU Delft, Supervisor, Committee chair
	Prof. Marius Staring,	Leiden University Medical Center, Supervisor
	Prerak Mody,	Leiden University Medical Center, Daily Supervisor
	Viktor van der Valk,	Leiden University Medical Center, Daily Supervisor
	Dr. Jie Yang,	TU Delft, External committee member

TU Delft

Preface

This thesis report presents my Master's Project work about analyzing Components of a Transformer under Different Data Scales in 3D Prostate CT Segmentation. The project was supervised by Dr. Jan van Gemert and Marius Staring. In addition, I received help from my daily supervisor Prerak Mody and Viktor van der Valk. The report is divided into three parts: The first part introduces the high-level storyline of the research; the second part is the scientific paper of the main work; the third part provides the technical explanations of core concepts included in part two.

I have learned a lot from my supervisors and had a pleasant journey throughout this process. Therefore, I would like to express my gratitude to all who have provided me with their help and patience with my silly mistakes during this process. I am grateful to Dr. Jan van Gemert and Prof. Marius Staring, who have instructed me to conduct proper research during their meetings. I can not miss the opportunity to express my gratitude to Prerak Mody and Viktor van der Valk who provided me with guidance when I was stuck. Furthermore, Mody inspired and helped me to submit my first paper. With his advice, I improve not only in research skills but also my communication with others. In addition, I am grateful to Dr.Jie Yang for spending his time as my committee member. The encouragement in his Information Retrieval course test means a lot to me. I especially appreciate the help and accompany from my girlfriend, Chengcheng Zhao as a non-expert reader.

Yicong Tan
Delft, September 2022

Contents

1
Introduction

Transformer[19], originally proposed for the machine translation task, has become the state-of-art methods for tasks in both Natural Language Processing[2, 7] and Vision domains[1, 4, 6, 8, 13, 20, 21]. Cancer treatment via radiotherapy requires the segmentation of tumors and organs-at-risk (OAR) on diagnostic scans like CT. Convolution-based UNet architectures [11] have dominated this task for years. However, the recent success of Transformers in the vision domain has raised the question whether they can replace convolutions as the primary image processing operation in medical segmentation [10, 18, 22].

1.1. Report Structure

The report is divided into three parts: the part 1 briefly introduces the motivation, research question and experiments; the part 2 is the scientific paper which describes the project in detail; The part 3 is the supplements materials to help readers understand the project.

The part 3 introduces vanilla Transformer[8], Swin-Transformer[14] and nnFormer[22] successively, which helps with the understanding of nnFormer, the window-based Transformer, used in part 3. A brief introduction of position encodings is included to get clearer awareness of experiments in part 2. Finally, dataset description in part 3 explains the organs at risk in the experiments by showing the ground truth visual masks of different organs.

1.2. Motivation

Cancer treatment via radiotherapy requires 3D CT segmentation to help with the dose distribution by segmenting tumors and organs at risk. Convolutions[11] have dominated the medical segmentation domain for years. However, the recent success of Transformers in the vision domain inspires the application of Transformers in 3D medical segmentation to provide more accurate segmenting masks. Preceding studies[9, 10, 18, 22] have applied the Transformer to 3D segmentation by replacing the convolutions with Transformer blocks in the UNet structure. However, due to the scarcity of medical data, one of the deficiencies of such work is that their data sets are small, leading to potential overfitting [9, 10, 22]. Consequently, the Transformer and convolution are compared when they both overfit on small data scales and the results may not be sufficient to differentiate them. Another deficiency is that Transformers and convolutions are compared in different general structures without ensuring equivalent parameter count[10, 17, 18, 22]. In addition, no existing method has closely examined the different components of the Transformer. The researchers[17, 18] who compare the Transformers with convolutions draw their conclusion by comparing two Transformer-based and convolution-based models but not investigating the reason behind. Inspired by ConvNeXt[16], we analyze different components of a transformer and replace them with more traditional deep learning operations like convolutions and pooling while ensuring equivalent parameter count and similar neural architectures. To solve the overfitting problem, we construct six different data scales to evaluate the Transformer and the compared model on both small and large data scales. In addition, unlike previous methods which train and test on a same dataset, we evaluate on a separate dataset with different clinical protocols for CT scan acquisition to test the generalization capability of the compared models.

1.3. Research Questions

Inspired by the Convnext[16] and driven by the above motivations, we attempt to answer our main research question:

- **What is the effect of various components of a Window-Based Transformer under different data scales?**

We answer this question by evaluating the performance of the Window-Based Transformer in context of replacing its various components in different data scales. As shown in figure 1.1, we aim to answer this main question by conducting three experiments successively.

It is often claimed that the Transformer has an advantage over the convolution in 3D medical segmentation. And Swin-Transformer block, the fundamental building block of the window-based Transformer, often serves as the replacement for convolutions. Therefore, we bring out our first question:

- What is the role of Swin-Transformer blocks compared to convolution blocks on different data scales?

The results show that Swin-Transformer blocks perform poorly in comparison with convolutions on large data scales. Surprisingly, we find that convolution also benefits more from both increased data scales and pretraining weights compared to the Transformer. In addition, though Transformer shows small advantage over convolution on small data scales, both of them suffer from overfitting. Therefore, we hypothesize that a simpler operation can outperform both Transformer and convolution on small data scales by reducing overfitting. In this case, we raise the question:

- Can a simpler operation Max-Pooling take the place of the Self-Attention mechanism on small data scales?

The results of the pooling operation suggest that the pooling outperforms both Transformer and convolution on small data scales. This indicates that it may always be better to choose simpler operations in low data regimes. We believe Transformer's underperformance could be explained by the lack of understanding of position information of voxels in our data scales. Therefore, we investigate different position encodings in the Transformer by asking the question: What is the effect of different position encodings?

- What are the effects of different position encodings?

We compare the absolute position embedding(learned and Sinusoid), relative position bias, and no position encodings in this experiment. The results indicate that the gaps in performance across different position encodings and no encodings are not large in all experiment settings. This further supports our hypothsis that the Transformer's underperformance could be explained by the lack of understanding in position information of voxels on our data scales.

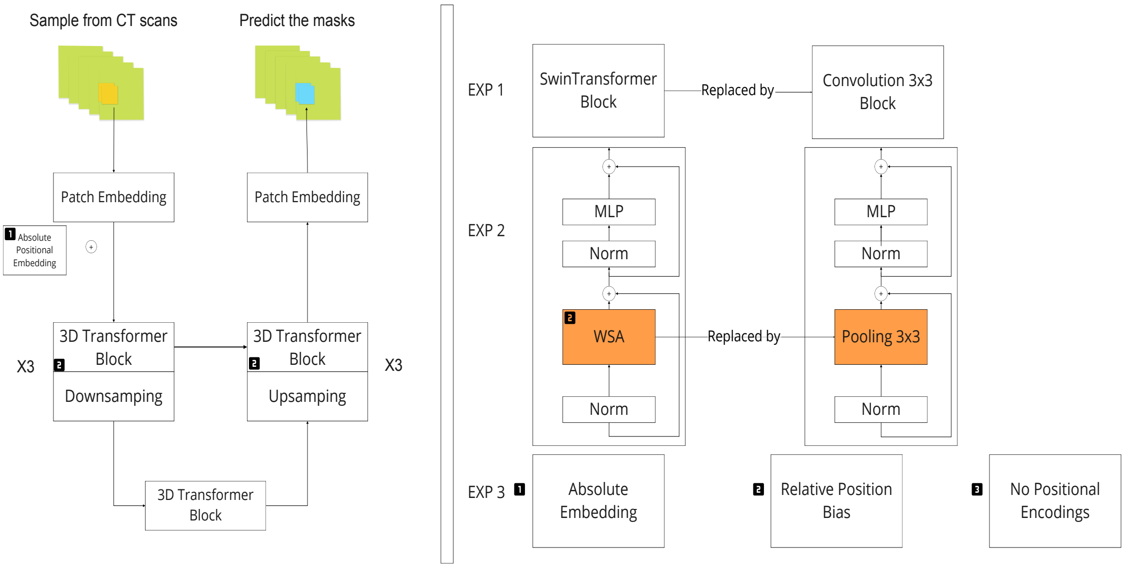


Figure 1.1: The visual abstract for our network architecture and our three experiment settings. On left part, the yellow tiles of squares denote the sampled sub-volumes from the CT scans; The blue tiles of squares denote the predicted masks for segmentation; The 3D Transformer blocks used are the 3D Swin-Transformer block. On the right part, we show the evaluated components and the replacement for each: The orange blocks in EXP denote that we only replace the Window-based Self-Attention with pooling. We use labels with black background to show the place to insert position encodings: the absolute position embeddings are added once to each position on the feature map after patch embedding; the relative bias is added in computing Self-Attention matrix in each Swin-Transformer block.

2
Scientific Paper

Analyzing Components of a Transformer under Different Data Scales in 3D Prostate CT Segmentation

Yicong Tan
Pattern Recognition Lab, TU Delft, Delft, The Netherlands
Y.Tan-2@student.tudelft.nl

September 22, 2022

Abstract

Literature on medical imaging segmentation claims that self-attention-based Transformer blocks perform better than convolution in UNet-based architectures. This recently touted success of Transformers warrants an investigation into which of its components contribute to its performance. Moreover, previous work has a limitation of analysis only at fixed data scales as well as unfair comparisons with others models where parameter counts are not equivalent. This work investigates the performance of the window-Based Transformer for prostate CT Organ-at-Risk (OAR) segmentation at different data scales in context of replacing its various components. To compare with previous literature, the first experiment replaces the window-based Transformer block with convolution. Results show that the convolution prevails as the data scale increases. In the second experiment, to reduce complexity, the self-attention mechanism is replaced with an equivalent albeit simpler spatial mixing operation i.e. max-pooling. We observe improved performance for max-pooling in smaller data scales, indicating that the window-based Transformer may not be the best choice in both small and larger data scales. Finally, since convolution has an inherent local inductive bias of positional information, we conduct a third experiment to imbibe such a property to the Transformer by exploring two kinds of positional encodings. The results show that there are insignificant improvements after adding positional encoding, indicating the Transformers deficiency in capturing positional information given our data scales. We hope that our approach can serve as a framework for others evaluating the utility of Transformers for their tasks. Code is available via `https://github.com/prerakmody/window-transformer-prostate-segmentation`.

1. Introduction

Transformer [42], originally proposed for the machine translation task, has become the state-of-art methods for tasks in both Natural Language Processing [2, 12] and Vision domains [1, 8, 11, 13, 29, 45, 51]. In vision domain, ViT [13] set the foundation by creating a general structure for applying the Transformer blocks in classification tasks. However, when it comes to extending the application to other tasks such as object detections and semantic segmentation, ViT suffered from certrain restrictions. The subsequent methods have broadened the use of Transformer by addressing the limitations in Vanilla ViT. For example, those methods have reduced the complexity in the Self-Attention mechanism [6,20,22,30], created multi-scale feature maps [30, 32, 44, 57] and introduced the inductive biases from convolutions [9, 11, 28, 30, 45, 49]. Encouraged by the previous work, researchers are attempting to apply the Transformer to the medical imaging domain.

Cancer treatment via radiotherapy requires the segmentation of tumors and organs-at-risk (OAR) on diagnostic scans like CT. Convolution-based UNet architectures [24] have dominated this task for years. However, the recent success of Transformers in the vision domain has raised the question whether they can replace convolutions as the primary image processing operation in deep learning [18, 38, 54]. In particular, are Transformers capable of replacing convolutions in 3D medical segmentation, where locality bias of convolutions plays an important part in segmenting borders between organs?

Specifically in the 3D medical segmentation, the window-based Transformer [54] has been used since the vanilla Transformer suffered from a computational complexity quadratic to the image size. Nevertheless, due to the scarcity of medical data, one of the deficiencies of such work is that their data sets are small, leading to potential overfitting [18,54]. Consequently, the Transformer and convolution are compared when they both overfit in small data scales. Our work remedies this by analysing transformers in a UNet-based architecture at six different data scales. In addition, we use a separate test set to evaluate the generalization capability for compared models. Another deficiency is that the Transformer and convolutions are compared in different general structure without ensuring equivalent parameter count [18, 54, 54]. Transformers usually have several times parameters when compared to convolution. For example, nnFormer with 150m parameters are compared to nnUNet with 30m parameters. Transformers also have different network architectures to the compared convolution method. The advantages may result from techniques such as skip connections, random paths, and Layer Normalization. Inspired by ConvNeXt [31] we analyze different components of a transformer and replace them with more traditional deep learning operations like convolutions and pooling while ensuring equivalent parameter count and similar neural architectures. Our results indicate that window-based Transformers perform worse than the comparison model in all our data scales for 3D prostate CT segmentation. Perhaps, Transformers need to evolve further to replace convolutions in the medical segmentation domain.

2. Related Work

In this section we first review UNet, which is the widely-used architecture in medical segmentation. Then, we go through the previous Transformer applications in the medical segmentation.

UNet [37], first applied in 2D slices and then extended to 3D CT scans and MRI images [10], is one of the fundamental convolution-based architectures in medical image segmentation. After the success of the initial UNet, several improved models that incorporated ideas from other domains based on the original UNet have emerged. For instance, the success of ResNet [19] and DenseNet [21] stimulated the development of ResUNet [47], ResUNet++ [25], Multi-ResUNet [23] and DenseUNet [16]. Additionally, the aggregation of the output from the deep and shallow layers inspired the UNet++ [55, 56]. Besides, the combination of attention mechanisms and UNet resulted in the Attention-UNet [35], Attention-UNet++ [27], MA-UNet [3], SCAU-Net [53], and AA-UNet [36].

Similar to the techniques above, the success of Vision Transformer motivated researchers to apply the Transformer blocks to the UNet structure. The Transformer blocks were used to extract the long-range dependencies in the image and were placed at deep layers due to computational complexity, while convolution blocks were used to extract low-level feature maps and were placed at shallow layers, such as UT-Net [14], TransUnet [7], MCTrans [26] and TransClaw U-Net [5]. In the meanwhile, transferring the architecture directly from the vision domain has become a trend, for example, Swin-Transformer [30] to Swin-UNet [4] and LeVit [15] to LeVit-UNet [48]. MedT [41] and MBT-Net [52] adopted the Axial-Attention [20] block to reduce the computational complexity while taking the advantage of the Self-Attention mechanism.

Researchers also apply the Transformer to 3D medical segmentation. However, this process is still at the initial stage. TransBTS [43] used the Transformer block to fuse the feature maps from 3D ConvNets. Besides, UNETR [18] replaced the convolution blocks with Transformer blocks in the 3D-UNet encoder. On top of that, Swin-UNETR [17] changed the vanilla Transformer blocks with Swin-Transformer blocks. Similarly, nnFormer [54] replaced convolutions with Swin-Transformer blocks in both the encoder and the decoder and incorporated the model in the frame of the nnUNet [24]. In addition, D-Former [46] was inspired by the dilated convolution and restricted the Self-Attention in a dilated block to reduce the complexity and enlarge the receptive field. We also built our experiment based on the nnUNet [24] that provides a general framework to handle arbitrary medical segmentation datasets by condensing and automating the segmentation pipeline. By doing so, we simplified the experiment's design of incorporating Transformers into the UNet.

Researchers compared convolutions with Transformers in the medical domain [33, 38]. The method introduced by Christos Matsoukas [33] was restricted to 2D segmentation. And both methods did not delve into the different components of the Transformer, nor did they create different data scales in comparison. To dig deeper, We were inspired by the Convnext [31] which replaced the components of a ResNet step by step and surpasses the performance of a Swin-Transformer. In addition, we have adopted the idea that the general architecture of the Transformer plays a significant role in performance from MetaFormer [50], MLP-Mixer [34, 39] and Conv-Mixer [40]. They splitted the Transformer block in to two parts: the Self-Attention correponds to the spatial-mixing and the feed-forward net goes with the channel-mixing. Both parts can be replaced by other existent deep-learning operations while maintaining the performance.

3. Method

3.1. Data

We use prostate CT data containing annotations of four organs: bladder, prostate, rectum, and seminal vesicles. The data is collected from three institutes, c.f. Haukeland Medical Center of Norway (HMC), Leiden University Medical Center in the Netherlands (LUMC) and Erasmus Medical Center in the Netherlands (EMC), containing 179, 475 and 56 CT scans, respectively. EMC is used as the test data set, while HMC and LUMC are used as the training datasets. Due to differences in clinical protocols for CT scan acquisition, the EMC dataset has larger volumes of the prostate and bladder, which makes it a challenging test dataset.

3.2. Network Architecture

As shown in Figure 1, the Window-based Transformer network in evaluation is nnFormer [54], which employs

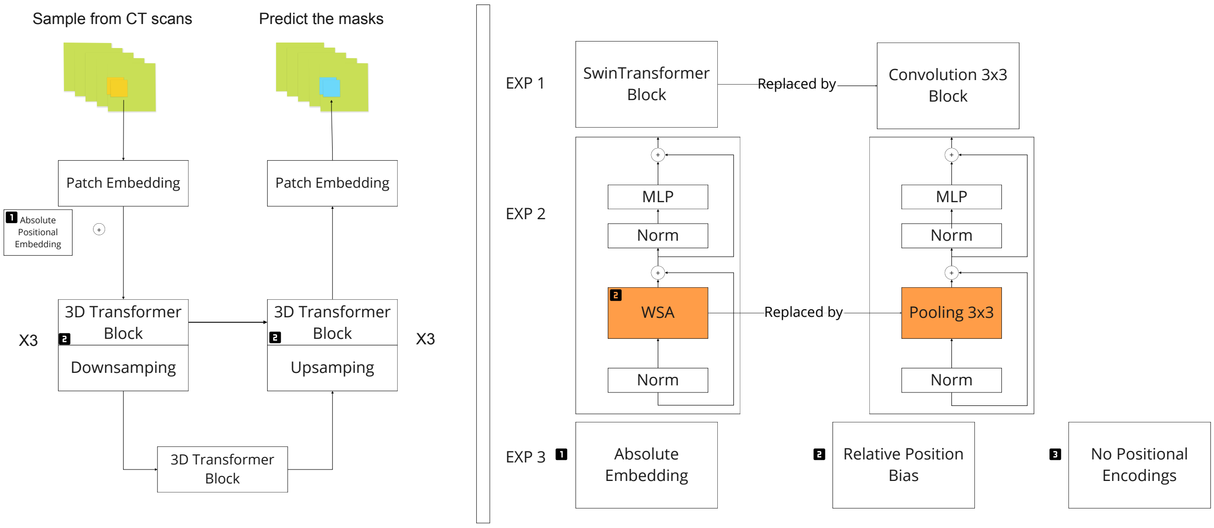


Figure 1: The visual abstract for our network architecture and our three experiment settings. On left part, the yellow tiles of squares denote the sampled sub-volumes from the CT scans; The blue tiles of squares denote the predicted masks for segmentation; The 3D Transformer blocks used are the 3D Swin-Transformer block. On the right part, we show the evaluated components and the replacement for each: The orange blocks in EXP denote that we only replace the Window-based Self-Attention with pooling. We use labels with black background to show the place to insert position encodings: the absolute position embeddings are added once to each position on the feature map after patch embedding; the relative bias is added in computing Self-Attention matrix in each Swin-Transformer block.

Swin (shifted-window)-Transformer blocks in the encoder and decoder of a UNet architecture. Please note that the first two layers of this architecture are convolution-based patch-embedding layers to extract low-level feature maps. Formula 1 shows the computation in a Window-based Transformer block for Layer l and l+1: $X^{l-1} \in \mathbb{R}^{B\times N\times L\times C}$ is the flatten output of the last layer, where B, N, L, C denote the batch size, the number of 3D windows, the number of tokens in one window, and the channel size, respectively. To flatten the feature map and divide it into multiple windows, we first split the feature map in to N parts where $N = \lceil \frac{H}{M} \rceil \times \lceil \frac{W}{M} \rceil \times \lceil \frac{D}{M} \rceil$, H, W, D, M denote the height, weight, depth of the CT scans and the 3D window size. Then we flatten each window by reshaping the 3D window to 1D sequence where the sequence length $L = M \times M \times M$.

$$\begin{aligned}
\hat{X}^l &= \mathrm{W\text{-}MSA}(\mathrm{LN}(\mathrm{X}^{l-1})) + \mathrm{X}^{l-1} \\
X^l &= \mathrm{MLP}(\mathrm{LN}(\hat{\mathrm{X}}^l)) + \hat{\mathrm{X}}^l \\
\hat{X}^{l+1} &= \mathrm{SW\text{-}MSA}(\mathrm{LN}(\mathrm{X}^l)) + \mathrm{X}^l \\
X^{l+1} &= \mathrm{MLP}(\mathrm{LN}(\hat{\mathrm{X}}^{l+1})) + \hat{\mathrm{X}}^{l+1}
\end{aligned} \tag{1}$$

The Window-based Multi-Head Self-Attention is then computed in each 3D window as shown in formula 2, in which Q, K, V are queries, keys and values, each linear transformed by the input flattened sequence; d denotes the size of the key and query. To compute Multi-Head Self-Attention in a shifted-window based, each 3D window is shifted towards right bottom by $(\lfloor \frac{M}{2} \rfloor, \lfloor \frac{M}{2} \rfloor, \lfloor \frac{M}{2} \rfloor)$ voxels. Besides, the bottom and rightmost voxels are shifted to the top and leftmost, since there is no extra space on the right bottom side.

$$\mathrm{Attention}(\mathrm{Q}, \mathrm{K}, \mathrm{V}) = \mathrm{Softmax}(\frac{\mathrm{QK}^{\mathrm{T}}}{\sqrt{\mathrm{d}}})\mathrm{V} \tag{2}$$

As shown in the left part of Figure 1, we have three experiment settings to examine the different components of the Transformer network: First, two Swin-Transformer blocks in one layer are replaced by two 3-by-3 convolutions; second, the Self-Attention mechanism within the block is replaced by the pooling operation; third, different position encodings are compared with the model without any position encodings.

3.2.1 Method 1: Replacing Swin-Transformer block with convolution block

Literature on medical image segmentation has shown superior performance of window-based Transformers over convolutions [18, 54]. We test this notion by replacing the window-based Transformer blocks with a sequence of two convolutions, as shown in formula 3. We also ensure that their parameter counts are equivalent and proceed to compare these models across multiple data scales. It is hypothesized that the Transformer will perform poorly in low-data regimes, since its attention mechanism is incapable of understanding relative position information of voxels, a quality important for precise tasks like segmentation and inherent to convolutions. Conversely, the lack of an inherent prior for imaging data, may allow Transformers to learn complex dependencies in the large-data regime, hence boosting performance.

$$\begin{aligned} X^l &= \mathrm{Conv}(\mathrm{LN}(\mathrm{X^{l-1}})) + \mathrm{X^{l-1}} \\ X^{l+1} &= \mathrm{Conv}(\mathrm{LN}(\mathrm{X^{l}})) + \mathrm{X^{l}} \end{aligned} \tag{3}$$

3.2.2 Method 2: Replacing the Self-Attention with Pooling

In the spirit of further analyzing components of the Transformer block and inspired by the MetaFormer [50] to reduce computational complexity, we replace the attention mechanism with a much simpler spatial feature mixing operation, i.e pooling, as shown in formula 4. Replacing the complex attention mechanism with a simpler pooling operation may also reduce the chance of overfitting in low-data regimes. We hypothesize that max-pooling will outperform self-attention in small data scales while self-attention will prevail gradually with increased data scale. This is because the complex nature of the attention mechanism when compared to max pooling might allow it to model spatial features provided additional data.

$$\begin{aligned} \hat{X}^l &= \text{Max-Pooling}(\mathrm{LN}(\mathrm{X^{l-1}})) + \mathrm{X^{l-1}} \\ X^l &= \mathrm{MLP}(\mathrm{LN}(\mathrm{\hat{X}^{l}})) + \mathrm{\hat{X}^{l}} \\ \hat{X}^{l+1} &= \text{Max-Pooling}(\mathrm{LN}(\mathrm{X^{l}})) + \mathrm{X^{l}} \\ X^{l+1} &= \mathrm{MLP}(\mathrm{LN}(\mathrm{\hat{X}^{l+1}})) + \mathrm{\hat{X}^{l+1}} \end{aligned} \tag{4}$$

3.2.3 Method 3: Evaluating Positional Encoding

Under the assumption that failures of window-based Transformers might be due to its inability to model positional dependencies, we explore two different positional encoding methods and compare them with a model without any positional encodings, as shown in formula 5. The first is absolute positional embedding that is added to the feature map after the convolutional patch-embedding. Therefore, the added absolute position embedding has the same dimension as the feature map input to the first Transformer block. Moreover, the absolute positional embedding can be divided into learned and unlearned positional embedding. We can extend original 1D sinusoid positional embedding to 3D case by following the formula 6. The second method is the relative positional bias that is added when computing the attention matrix in each Swin-Transformer block. Our base Transformer model uses relative positional bias which we expect to perform better as per work done in literature [54].

$$\begin{aligned} X^0 &= X^0 + Z^{embedding} \\ \mathrm{Attention(Q,K,V)} &= \mathrm{Softmax}(\frac{\mathrm{QK^T}}{\sqrt{\mathrm{d}}} + \mathrm{B})\mathrm{V} \end{aligned} \tag{5}$$

$$\begin{aligned} &\mathrm{PE(x,y,z,2i) = sin(x/10000^{6i/D})} \\ &\mathrm{PE(x,y,z,2i+1) = cos(x/10000^{6i/D})} \\ &\mathrm{PE(x,y,z,2j+D/3) = sin(y/10000^{6i/D}))} \\ &\mathrm{PE(x,y,z,2j+1+D/3) = cos(y/10000^{6i/D})} \\ &\mathrm{PE(x,y,z,2k+2D/3) = sin(z/10000^{6i/D})} \\ &\mathrm{PE(x,y,z,2k+1+2D/3) = cos(z/10000^{6i/D})} \end{aligned} \tag{6}$$

4. EXPERIMENTS AND RESULTS

4.1. Experiment Settings

This work uses two datasets for training i.e. HMC (or clinic A) and LUMC (or clinic B). The HMC dataset is split into two parts for 2-fold cross validation and also creating smaller data scales. They contain 94 and 85 CT scans respectively and are henceforth referred to as A1 and A2. We make 6 combinations of these datasets c.f. A1, A2, A, A1+B, A2+B, A+B to create multiple data scales. Note, that the data from clinic B is not used for pretraining, but rather as additional scans during training. In addition, we use the clinic B to pretrain the models and then finetune on A1, A2, and A so as to test the compared models' performance in pretraining context. Three experiments are conducted to compare the window-based Transformer to its counterparts on two geometric metrics i.e. Dice and 95$^{\mathrm{th}}$ percentile Hausdorff Distance (HD95) averaged over all scans of the test dataset. In addition, we adopt wilcoxon signed rank test with p value equal to 0.5 to reveal the statistical significance between two compared models in each organ and average of all organs.

The models are trained using a combination of Dice and cross-entropy loss in deep supervision for 500 epochs.

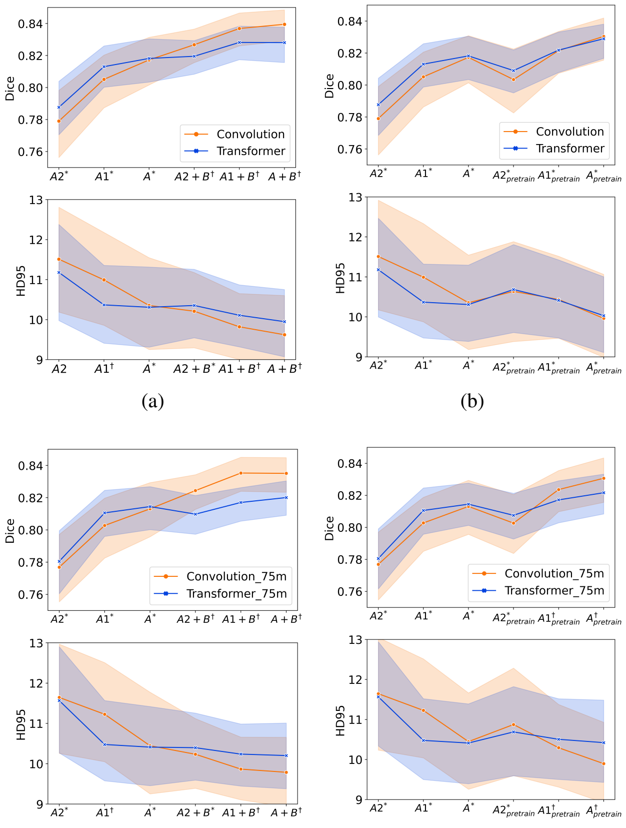
The window-based Transformer and convolution contain 158.49M and 155.85M parameters, respectively. The CT scans are first resampled to the median spacing of each dataset and then randomly sampled patches of size (128,128,64) are the input to the network. Models were trained with Pytorch 11.3 on a single Nvidia RTX6000 (24GB memory). In addition to the lineplots in Figure 2, the full test experiment results are in Appendix A.

4.2. Experiment 1: Replacing Swin-Transformer block with convolution block

Surprisingly, Figure 2 (a) and Figure 6 (a)(b)(c)(d) show that the Transformer performs better on lower data scales and the convolution gradually surpasses it with the increase of data. The convolution performs poorly in the lower-data regime since the lack of data coupled with its locality bias may not allow it to learn sufficient global shape-based information, but only local textural information in the neighbourhood of a voxel (*seminal vesicle in Figure 6 (b)*). A lower supervision loss during training and higher performance in cross-validation experiments on clinic A also are indicators of the overfitting nature of convolutions in our smaller data regimes. The higher performance of convolutions in our larger data scales may imply that the Transformer needs more data to learn the dependencies within the data (*jagged nature of bladder in Figure 6 (c)*).

In the pretraining context, Figure 2 (b) and Figure 6 (i)(j)(k)(l) suggest that both the Transformer and the convolution benefit from pretraining. However, compared to the results on small data scales, the gaps between the Transformer and convolution are eliminated after pretraining. In this light, the convolution benefits more from pretraining compared to the Transformer. Nevertheless, Figure 6 (i)(j)(k)(l) show that both pretrained models still have insufficient predictions compared to the ground truth, especially the zigzag border lines in 6 (i),(k) due to the lack of locality bias in the Transformer block.

Next, we conduct additional experiments to further verify our conclusions. We halve the parameters both for the Transformer and convolution models to 75m and repeat the experiment. Figure 2 (c)(d) shows that after we halved the parameters, the Transformer performs better on lower data scales and the convolution gradually surpasses it with increased data or within pretraining context. It is identical to the experiments in the original parameter count. In addition, compared to the initial experiments, the performance gap enlarges between the Transformer model and the convolution model in large data scales and pretraining context. This indicates that the Transformer is more sensitive to the parameter reduction compared to the convolution in our experiment settings.



(a)

(b)

(c)

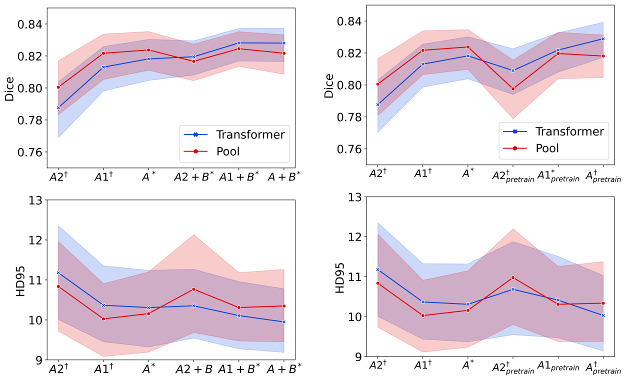
(d)

Figure 2: Experiments 1: Line plots showing the mean and 95th percentile confidence interval of Dice and HD95. The x-axis denotes the various training data scales with clinics A and B while A2, A1 denote the first and second part of clinic A. The subscript Pretrain denotes that the model is pretrained on B and finetuned on given dataset. $*$ denotes a statistical difference in at least one organ and † denotes a statistical difference on average and for at least one organ.

4.3. Experiment 2: Replacing the Self-Attention with Pooling

In line with our expectations, Figure 2 (b) and Figure 6 (e)(f)(g)(h) suggest that max-pooling outperforms in small data scales compared to the window-based self-attention mechanism and the latter surpasses with the increase of data scale. Thus, both convolution and window-based Transformer fail to be well-trained under small data scales. These results indicate that the simplicity of pooling may be essential to high performance in small data regimes.

Apart from that, Figure 3 (a)(b) and Figure 6 (m),(n),(o),(p) show that the Transformer model benefits from pretraining, while the Pooling operation fails to elevate the performance even with pretraining. This might fur-



(a)

(b)

Figure 3: Experiments 2: Line plots showing the mean and 95^{th} percentile confidence interval of Dice and HD95. The x-axis denotes the various training data scales with clinics A and B while A2, A1 denote the first and second part of clinic A. The subscript Pretrain denotes that the model is pretrained on B and finetuned on given dataset. $*$ denotes a statistical difference in at least one organ and † denotes a statistical difference on average and for at least one organ.

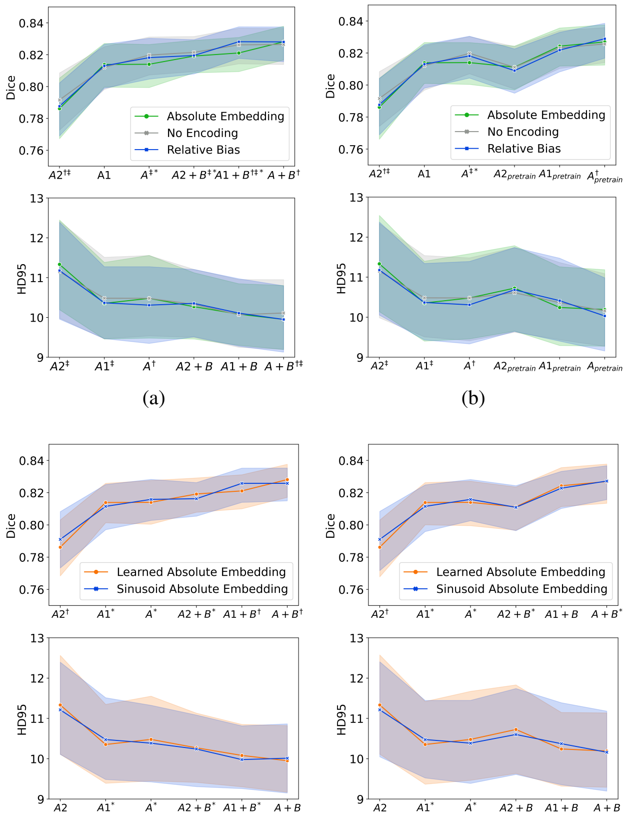
ther indicate that a larger data scale compared to our clinic A is desired for our Transformer model.

In addition, we have compared the average-pooling and max-pooling in the smallest three data scales. Figure 5 (d) indicates that max-pooling has an advantage over the average-pooling. A possible explanation for this is that the max-pooling extracts the important local features, which contributes to segmenting the edges and borders, while the average-pooling smooths them.

Similar to the first experiment, pooling also outperforms in the seminal vesicle on small data scales

4.4. Experiment 3: Evaluating Positional Encoding

Figure 4 (a),(b),(c),(d) shows that in spite of statistical differences, the gaps in performance across the different positional encodings is not large in all experiment settings. The lack of a large difference between the models with some form of positional encoding and those without, indicates that the current data scales are either insufficient to train the positional encodings well or that a better positional encoding design is needed for medical segmentation. Contrary to Transformers, both convolutions and pooling have some form of inductive bias (i.e. locality and neighbourhood structure). This could be one reason that the window-based Transformer is not the best choice in both our smaller and larger data scales.



(a)

(b)

(c)

(d)

Figure 4: Experiment 3: Line plots showing the mean and 95^{th} percentile confidence interval of Dice and HD95. The x-axis denotes the various training data scales with clinics A and B while A2, A1 denote the first and second part of clinic A. The subscript Pretrain denotes that the model is pretrained on B and finetuned on given dataset. In (a)(b), †, ‡ and $*$ denote a statistical difference between the relative bias and no encoding, absolute embedding and no encoding, and the two positional encoding respectively on the average of all organs. In (c)(d), $*$ denotes a statistical difference in at least one organ and † denotes a statistical difference on average and for at least one organ.

4.5. Side Experiments

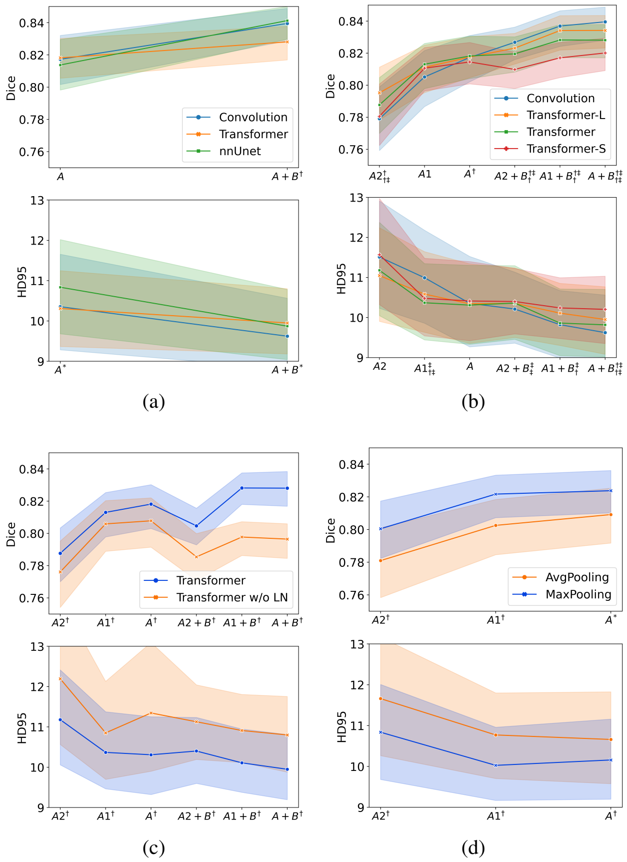
An additional side experiment is conducted to compare the performance of nnUnet and the convolution and Transformer models. Figure 5 (a) shows that nnUnet needs more data to surpass the Transformer. In this light, it often serves as a baseline in comparison with the Transformer models on data scales that are insufficient to meet its data hunger. In addition, we constructed the Transformer-S and Transformer-L by halving and doubling the parameters, respectively and compare them with both the Transformer and convolution. Figure 5 (b) shows that the convolution even outperforms the Transformer-L in the large data scales. Independent from the other experiments, we conduct the experiment to verify the effectiveness of the Layer Normalization. Figure 5 (c) shows that removing the Layer Normalization from the Transformer model decreases the performance and more in large data scales.

4.6. Validation Results

We conduct the two-fold cross-validation on dataset clinic B. As the validation set, clinic B shares more similarities with the training set compared to the test set clinic C, which is taken by different CT machines and annotated by distinct physicians. The results in the appendix B, show that convolution outperforms the Transformer in most small and large data scales(pretraining and finetuning); max-pooling performs equally with the Transformer in most data scales. This shows the convolution's more powerful capability in fitting the training data and predicting the validation data on current data scales. Besides, in line with the test set, the gaps in performance across the different positional encodings are not significant in all validation settings.

5. Discussion And Conclusion

This study evaluates different components of the window-based Transformer to understand their role in its performance. Unlike previous work, we maintain a constant parameter count across models and also analyze the effect of the components under different data scales and pretraining context. Our results show that window-based Transformers perform poorly in comparison with convolutions on large data scales. Surprisingly, we find that convolution also benefits more from both increased data scales and pretraining weights compared to the Transformer. We believe this underperformance could be explained by the lack of understanding of positional information of voxels in our data scales. The results of the pooling operation suggest that it may always be better to choose simpler operations in low data regimes. The comparable performance of models with and without positional encodings further supports our first claim. Thus, we conclude that for our dataset the window-based Transformer is not the best choice in both small and larger data scales. Please note that our largest data scale may not be sufficient for Transformers which are well-known to be data hungry. Future work could use our approach to understand Transformers by either segmenting other organs or using different medical imaging modalities. Additionally, it may be worth exploring self-supervised methods as they could potentially benefit the data hungry Transformer.



(a)

(b)

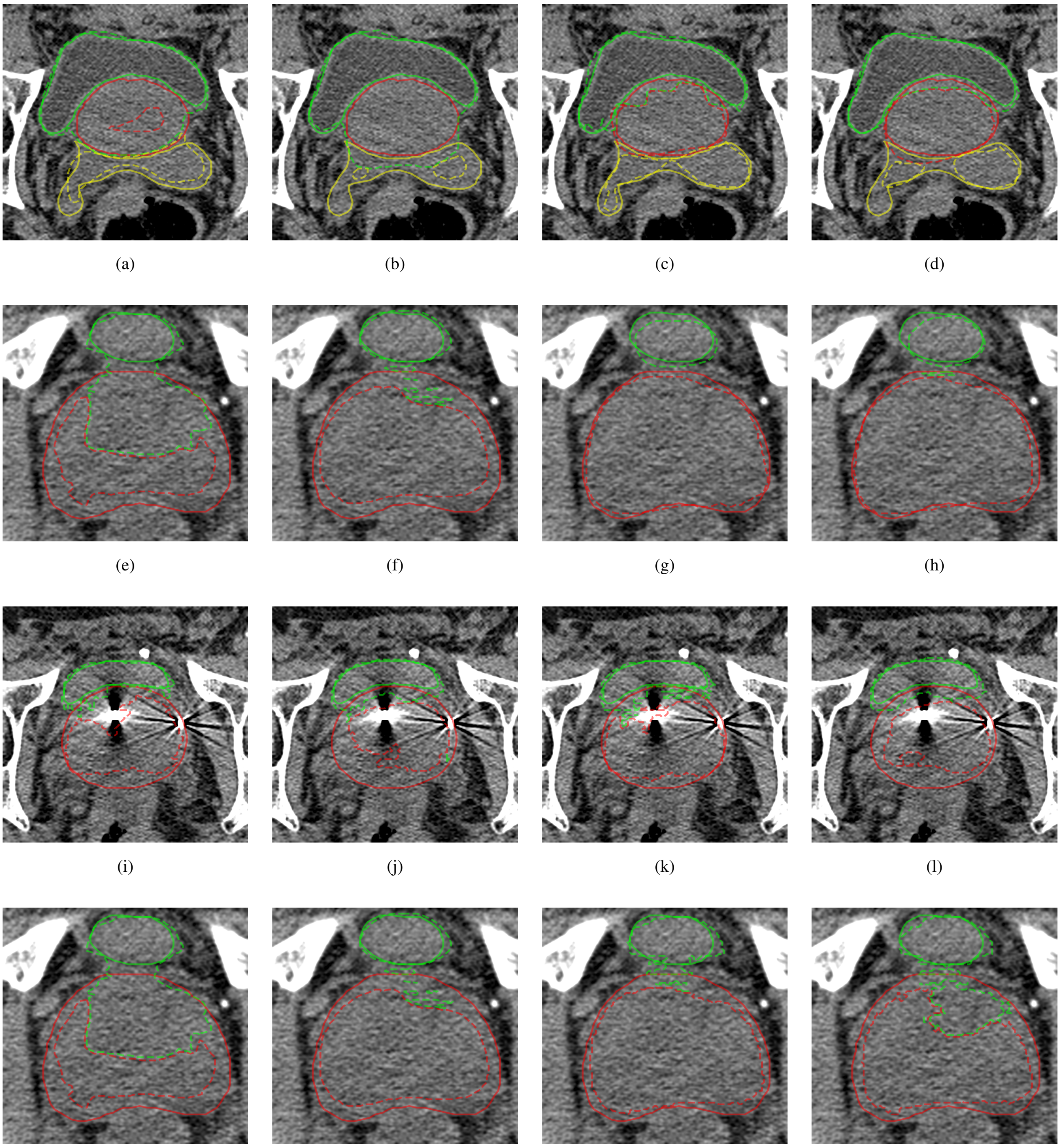
(c)

(d)

Figure 5: Side Experiments: Line plots showing the mean and 95$^{\text{th}}$ percentile confidence interval of Dice and HD95. The x-axis denotes the various training data scales with clinics A and B while A2, A1 denote the first and second part of clinic A. In (a), †, ‡ and $*$ denote the statistical difference between the relative bias and no encoding, absolute embedding and no encoding, and the two positional encoding respectively on the average of all organs. In (b), The subscript and superscript denote the statistical difference of the Transformer-L and Transformer-S respectively; †, ‡ denote a statistical difference with the Transformer and convolution model respectively. In (c)(d), †denotes a statistical difference on average and for at least one organ.

Acknowledgement

I do appreciate the useful advice and help from Professor Gemert and Staring, daily supervisor Mody and Viktor. In addition, I would like to express the my gratitude to my girl friend Chengcheng to her help as a non-expert reader.



(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j) (k) (l)

(m) (n) (o) (p)

Figure 6: CT scans showing the prediction (dotted line) and ground truth (solid line) for the prostate (red), bladder (green) and seminal vesicle (yellow). (a),(b),(e),(f) and (c),(d),(g),(h) show results when trained on the smallest and largest data scale respectively. (i),(j),(m),(n) and (k),(l),(o),(p) show the results when trained on the smallest and with Pretrained-Finetuned on largest data scale respectively

References

[1] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 1

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1

[3] Yutong Cai and Yong Wang. Ma-unet: An improved version of unet based on multi-scale and attention mechanism for medical image segmentation. In *Third International Conference on Electronics and Communication; Network and Computer Technology (ECNCT 2021)*, volume 12167, pages 205–211. SPIE, 2022. 2

[4] 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. 2

[5] Yao Chang, Hu Menghan, Zhai Guangtao, and Zhang Xiao-Ping. Transclaw u-net: Claw u-net with transformers for medical image segmentation. *arXiv preprint arXiv:2107.05188*, 2021. 2

[6] Chun-Fu Chen, Rameswar Panda, and Quanfu Fan. Regionvit: Regional-to-local attention for vision transformers. *arXiv preprint arXiv:2106.02689*, 2021. 1

[7] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021. 2

[8] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534*, 2022. 1

[9] Xiangxiang Chu, Zhi Tian, Bo Zhang, Xinlong Wang, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Conditional positional encodings for vision transformers. *arXiv preprint arXiv:2102.10882*, 2021. 1

[10] 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. 2

[11] Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. *Advances in Neural Information Processing Systems*, 34:3965–3977, 2021. 1

[12] 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. 1

[13] 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. 1

[14] 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. 2

[15] Benjamin Graham, Alaaeldin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin, Hervé Jégou, and Matthijs Douze. Levit: a vision transformer in convnet's clothing for faster inference. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12259–12269, 2021. 2

[16] Steven Guan, Amir A Khan, Siddhartha Sikdar, and Parag V Chitnis. Fully dense unet for 2-d sparse photoacoustic tomography artifact removal. *IEEE journal of biomedical and health informatics*, 24(2):568–576, 2019. 2

[17] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI Brainlesion Workshop*, pages 272–284. Springer, 2022. 2

[18] 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*, 2022. 1, 2, 4

[19] 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. 2

[20] Jonathan Ho, Nal Kalchbrenner, Dirk Weissenborn, and Tim Salimans. Axial attention in multidimensional transformers. *arXiv preprint arXiv:1912.12180*, 2019. 1, 2

[21] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 2

[22] Zilong Huang, Youcheng Ben, Guozhong Luo, Pei Cheng, Gang Yu, and Bin Fu. Shuffle transformer: Rethinking spatial shuffle for vision transformer. *arXiv preprint arXiv:2106.03650*, 2021. 1

[23] Nabil Ibtehaz and M Sohel Rahman. Multiresunet: Rethinking the u-net architecture for multimodal biomedical image segmentation. *Neural networks*, 121:74–87, 2020. 2

[24] 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*, 2021. 1, 2

[25] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Dag Johansen, Thomas De Lange, Pål Halvorsen, and Håvard D Johansen. Resunet++: An advanced architecture for medical image segmentation. In *2019 IEEE International Symposium on Multimedia (ISM)*, pages 225–2255. IEEE, 2019. 2

[26] Yuanfeng Ji, Ruimao Zhang, Huijie Wang, Zhen Li, Lingyun Wu, Shaoting Zhang, and Ping Luo. Multi-compound trans-

former for accurate biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 326–336. Springer, 2021. 2

[27] Chen Li, Yusong Tan, Wei Chen, Xin Luo, Yuanming Gao, Xiaogang Jia, and Zhiying Wang. Attention unet++: A nested attention-aware u-net for liver ct image segmentation. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 345–349. IEEE, 2020. 2

[28] Yun Liu, Guolei Sun, Yu Qiu, Le Zhang, Ajad Chhatkuli, and Luc Van Gool. Transformer in convolutional neural networks. *arXiv preprint arXiv:2106.03180*, 2021. 1

[29] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12009–12019, 2022. 1

[30] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 1, 2

[31] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2

[32] Teli Ma, Mingyuan Mao, Honghui Zheng, Peng Gao, Xiaodi Wang, Shumin Han, Errui Ding, Baochang Zhang, and David Doermann. Oriented object detection with transformer. *arXiv preprint arXiv:2106.03146*, 2021. 1

[33] Christos Matsoukas, Johan Fredin Haslum, Magnus Söderberg, and Kevin Smith. Is it time to replace cnns with transformers for medical images? *arXiv preprint arXiv:2108.09038*, 2021. 2

[34] Luke Melas-Kyriazi. Do you even need attention? a stack of feed-forward layers does surprisingly well on imagenet. *arXiv preprint arXiv:2105.02723*, 2021. 2

[35] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018. 2

[36] Kumar T Rajamani, Priya Rani, Hanna Siebert, Rajkumar ElagiriRamalingam, and Mattias P Heinrich. Attention-augmented u-net (aa-u-net) for semantic segmentation. *Signal, image and video processing*, pages 1–9, 2022. 2

[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. 2

[38] Ikboljon Sobirov, Otabek Nazarov, Hussain Alasmawi, and Mohammad Yaqub. Automatic segmentation of head and neck tumor: How powerful transformers are? In *Medical Imaging with Deep Learning*, 2022. 1, 2

[39] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in Neural Information Processing Systems*, 34:24261–24272, 2021. 2

[40] Asher Trockman and J Zico Kolter. Patches are all you need? *arXiv preprint arXiv:2201.09792*, 2022. 2

[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. 2

[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. 1

[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. 2

[44] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 568–578, 2021. 1

[45] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22–31, 2021. 1

[46] Yixuan Wu, Kuanlun Liao, Jintai Chen, Danny Z Chen, Jinhong Wang, Honghao Gao, and Jian Wu. D-former: A u-shaped dilated transformer for 3d medical image segmentation. *arXiv preprint arXiv:2201.00462*, 2022. 2

[47] Xiao Xiao, Shen Lian, Zhiming Luo, and Shaozi Li. Weighted res-unet for high-quality retina vessel segmentation. In *2018 9th international conference on information technology in medicine and education (ITME)*, pages 327–331. IEEE, 2018. 2

[48] Guoping Xu, Xingrong Wu, Xuan Zhang, and Xinwei He. Levit-unet: Make faster encoders with transformer for medical image segmentation. *arXiv preprint arXiv:2107.08623*, 2021. 2

[49] Yufei Xu, Qiming Zhang, Jing Zhang, and Dacheng Tao. Vitae: Vision transformer advanced by exploring intrinsic inductive bias. *Advances in Neural Information Processing Systems*, 34:28522–28535, 2021. 1

[50] Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2, 4

[51] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. 1

[52] Yinglin Zhang, Risa Higashita, Huazhu Fu, Yanwu Xu, Yang Zhang, Haofeng Liu, Jian Zhang, and Jiang Liu. A multi-branch hybrid transformer network for corneal endothelial cell segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 99–108. Springer, 2021. 2

[53] Peng Zhao, Jindi Zhang, Weijia Fang, and Shuiguang Deng. Scau-net: spatial-channel attention u-net for gland segmentation. *Frontiers in Bioengineering and Biotechnology*, 8:670, 2020. 2

[54] 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. 1, 2, 4

[55] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming 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*, pages 3–11. Springer, 2018. 2

[56] 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. 2

[57] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 1

A. Test Experiment results

The full experiment result is shown in this section. We found that Seminal Vesicle is a good indicator of performance since it is small and irregular compared to other organs.

A.1. Experiment1: Replacing Swin-Transformer Block with Convolution

The test experiment results for replacing Swin-Transformer Block with Convolution are shown in table 1 and table 2. The results after halving the parameters size are shown in table 3 and table 4.

		Mean		Prostate		Bladder		Rectum		SeminalVesicle	
Network	Training Set	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median
Trans	A	81.82+5.07	83.15	83.83+4.43	85.44	95.65+3.24	96.66	74.31+9.66	74.48	73.48+15.59	77.45
Conv	A	81.73+5.77	83.53	83.57+6.29	85.81	95.45+3.79	96.77	75.75+7.45†	76.09	72.14+17.07	78.27
Trans	A1	81.31+5.12	82.31	83.44+4.25	84.11	95.69+3.4	96.75	72.8+10.81	72.54	73.3+14.29	76.77
Conv	A1	80.51+6.55	82.08	81.46+9.95	84.46	95.8+3.57†	96.99	73.81+7.63	73.48	70.97+17.63†	75.56
Trans	A2	78.77+6.61	80.32	80.21+7.97	83.04	94.23+3.92	95.59	73.94+8.75	74.44	66.69+18.93	74.68
Conv	A2	77.9+7.8	80.15	80.84+8.94†	84.07	94.01+4.55	95.77	75.02+7.7†	74.37	61.75+20.71	69.37
Trans	A+B	82.81+3.97	83.28	84.56+5.33	86.08	95.58+2.56	96.26	73.29+7.5	73.26	77.81+9.96	79.05
Conv	A+B	83.95+3.88†	84.57	86.59+4.57†	87.72	96.13+1.85†	96.55	74.55+7.52†	74.39	78.55+10.48	80.39
Trans	A1+B	82.82+3.94	83.41	84.53+5.69	86.24	96.08+2.41	96.62	73.24+7.61	72.68	77.43+10.27	78.48
Conv	A1+B	83.69+3.95†	84.36	85.6+5.11†	87.11	96.58+1.72†	96.96	74.06+7.59†	73.47	78.52+10.45	79.17
Trans	A2+B	81.95+4.19	82.91	84.88+5.08	86.53	95.29+2.3	95.83	72.8+8.02	72.95	74.84+10.83	75.14
Conv	A2+B	82.68+4.07†	83.34	86.28+4.47†	87.54	95.64+2.21†	95.98	73.94+7.49†	73.75	74.84+10.86	76.26
Trans⋆	A	82.89+4.14	84.07	84.92+4.31	86.47	95.7+3.25	96.55	74.87+7.18	74.29	76.08+12.96	78.68
Conv⋆	A	83.04+4.94	84.34	85.16+5.41	87.42	95.8+3.2†	96.62	75.39+7.43	75.75	75.8+14.28	80.66
Trans⋆	A1	82.18+4.69	83.44	84.27+4.1	85.36	95.93+3.37	97.1	74.13+7.66	73.25	74.41+14.85	78.59
Conv⋆	A1	82.15+4.88	83.38	84.23+6.12	86.34	96.31+2.91†	97.14	74.22+7.52	73.34	73.86+14.62	77.22
Trans⋆	A2	80.9+5.39	82.78	83.02+5.6	85.09	94.69+3.69	95.78	74.45+7.53	74.83	71.45+16.22	75.86
Conv⋆	A2	80.34+7.3	82.85	83.79+7.2†	86.19	94.62+3.78	95.54	75.01+7.88†	74.9	67.94+19.47†	74.48

Table 1: Experiment 1: Replacing Swin-Transformer block with convolution block. The Dice mean and standard deviation on EMC Test set(clinic C). ⋆ denotes the pretraining on clinic B and finetuning on clinic A. † represents statistical significant difference(p<0.05 between the annotated method and Trans with the same training set

		Mean		Prostate		Bladder		Rectum		SeminalVesicle	
Network	Training Set	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median
Trans	A	10.31+3.77	9.46	5.42+2.38	4.5	2.98+2.57	1.89	26.96+11.64	26.61	5.88+4.39	4.49
Conv	A	10.35+4.51	9.07	5.84+3.8	4.54	3.7+4.59	1.81	26.42+11.79†	25.32	5.45+4.06	3.71
Trans	A1	10.37+3.59	9.5	5.5+2.25	4.87	3.2+2.76	1.89	27.17+10.63	25.91	5.61+3.84	4.64
Conv	A1	10.99+4.52†	9.35	6.45+4.56	5.06	3.35+3.35	1.76	28.06+11.32†	27.0	6.11+4.83	4.67
Trans	A2	11.18+4.56	9.99	6.58+3.54	5.73	4.31+3.94	2.59	26.93+11.96	26.71	6.89+4.94	4.92
Conv	A2	11.51+5.27	10.09	6.6+4.26	5.2	4.66+4.96	2.41	26.7+12.41	27.0	8.08+5.86	5.97
Trans	A+B	9.95+3.13	9.54	5.11+1.56	4.69	2.42+1.65	1.83	27.82+11.24	27.0	4.45+2.21	4.09
Conv	A+B	9.62+3.33†	9.12	4.65+1.42†	4.5	2.14+1.09†	1.75	27.39+12.02†	25.96	4.31+2.32	3.92
Trans	A1+B	10.11+3.1	9.53	5.2+1.59	4.88	2.3+1.76	1.79	28.32+11.15	27.27	4.61+2.44	4.11
Conv	A1+B	9.82+3.15†	9.32	4.94+1.4†	4.71	2.08+1.0	1.72	28.05+11.65	27.0	4.21+2.14	3.9
Trans	A2+B	10.35+3.31	9.89	5.09+1.5	4.58	2.44+1.41	1.99	28.55+11.98	28.26	5.33+2.64	4.64
Conv	A2+B	10.21+3.4	9.71	4.85+1.56†	4.51	2.33+1.42†	1.88	27.9+12.12†	27.01	5.75+2.93	5.69
Trans⋆	A	10.03+3.54	9.2	5.17+2.39	4.4	2.74+2.58	1.76	27.33+11.69	26.74	4.89+3.09	3.94
Conv⋆	A	9.96+4.04	8.87	5.18+3.08	4.48	2.97+3.4	1.79	26.86+11.71†	24.88	4.84+3.68	3.74
Trans⋆	A1	10.41+3.84	9.38	5.45+2.53	4.5	3.29+3.87	1.77	27.44+10.91	26.66	5.49+3.61	4.66
Conv⋆	A1	10.44+3.9	9.18	5.48+3.2	4.5	2.84+3.46†	1.73	27.93+11.43	26.92	5.49+3.68	4.55
Trans⋆	A2	10.68+4.16	9.75	5.74+3.16	4.67	3.46+3.13	2.23	27.24+11.85†	26.87	6.3+4.96	4.82
Conv⋆	A2	10.64+4.85	9.23	5.63+3.61	4.5	3.62+3.9	2.22	26.5+12.38	25.96	6.81+5.23	4.56

Table 2: Experiment 1: Replacing Swin-Transformer block with convolution block. The HD95 mean and standard deviation on EMC Test set(clinic C). ⋆ denotes the pretraining on clinic B and finetuning on clinic A. † represents statistical significant difference(p<0.05 between the annotated method and Trans with the same training set

		Mean		Prostate		Bladder		Rectum		SeminalVesicle	
Network	Training Set	$\mu+\sigma$	median	$\mu+\sigma$	median	$\mu+\sigma$	median	$\mu+\sigma$	median	$\mu+\sigma$	median
Trans	A	81.45+5.15	82.68	83.59+4.76	85.24	95.57+3.18	96.6	73.56+9.57	73.24	73.08+15.71	78.35
Conv	A	81.31+6.34	82.92	83.03+7.44	85.32	95.32+3.78	96.57	75.46+7.28†	75.48	71.43+17.75†	78.48
Trans	A1	81.06+5.38	82.59	83.17+4.39	84.14	95.68+3.37	96.93	72.8+10.94	72.57	72.58+15.0	77.05
Conv	A1	80.28+6.93	82.41	81.34+10.32	84.42	95.45+4.14	97.04	73.61+7.54	73.1	70.7+17.48†	75.82
Trans	A2	78.06+7.47	80.05	80.15+8.53	83.36	94.21+4.08	95.67	73.65+9.27	73.69	64.21+21.09	72.49
Conv	A2	77.69+8.24	80.17	79.58+10.87	83.61	93.72+4.9†	95.62	74.96+7.66†	74.26	62.5+20.53	69.99
Trans	A+B	82.01+4.02	82.77	83.98+5.68	84.6	95.41+2.54	96.07	72.17+7.79	71.94	76.47+9.57	77.56
Conv	A+B	83.51+3.93†	84.39	85.98+4.63†	87.42	95.97+2.08†	96.45	73.87+7.65†	73.59	78.22+10.41†	79.98
Trans	A1+B	81.7+4.1	82.15	83.43+6.02	85.09	95.51+2.48	96.01	71.85+7.52	71.07	76.02+10.21	77.88
Conv	A1+B	83.53+4.07†	84.11	85.42+5.08†	87.06	96.32+2.08†	96.79	73.87+7.56†	73.65	78.5+10.94†	79.71
Trans	A2+B	80.98+4.57	81.73	84.02+5.5	84.9	95.2+2.35	95.68	72.19+8.09	72.43	72.5+12.38	74.91
Conv	A2+B	82.44+4.12†	82.81	86.34+4.33†	87.27	95.47+2.11†	95.85	73.65+7.74†	73.23	74.3+11.06†	75.53
Trans	A2+B	80.98+4.57	81.73	84.02+5.5	84.9	95.2+2.35	95.68	72.19+8.09	72.43	72.5+12.38	74.91
Conv	A2+B	82.44+4.12†	82.81	86.34+4.33†	87.27	95.47+2.11†	95.85	73.65+7.74†	73.23	74.3+11.06†	75.53
Trans⋆	A	82.16+4.8	83.33	84.13+4.38	85.77	95.63+3.19	96.51	74.33+7.36	73.86	74.57+14.7	78.74
Conv⋆	A	83.07+5.09†	84.09	85.04+6.37†	87.26	95.86+3.07†	96.73	75.37+7.39†	75.56	76.0+13.71	79.61
Trans⋆	A1	81.72+4.97	83.26	83.87+4.27	85.05	95.86+3.46	97.07	73.6+7.93	72.63	73.54+15.38	78.12
Conv⋆	A1	82.36+5.05†	83.56	84.03+7.9†	86.71	96.27+3.13†	97.18	74.04+7.5	73.72	75.09+13.72	78.9
Trans⋆	A2	80.75+5.64	82.55	82.99+5.28	84.74	94.65+3.59	95.66	74.28+7.84	74.85	71.09+16.59	75.8
Conv⋆	A2	80.27+7.06	82.46	83.5+7.03†	85.92	94.59+3.86	95.64	74.58+7.89†	74.15	68.4+18.83	75.11

Table 3: Experiment 1: Replacing Swin-Transformer block with convolution block after halving the parameters. The Dice mean and standard deviation on EMC Test set(clinic C). ⋆ denotes the pretraining on clinic B and finetuning on clinic A. † represents statistical significant difference($p<0.05$ between the annotated method and Trans with the same training set

		Mean		Prostate		Bladder		Rectum		SeminalVesicle	
Network	Training Set	$\mu+\sigma$	median	$\mu+\sigma$	median	$\mu+\sigma$	median	$\mu+\sigma$	median	$\mu+\sigma$	median
Trans	A	10.41+3.84	9.76	5.47+2.48	4.76	3.09+2.63	1.95	27.43+11.68	26.86	5.66+4.02	3.91
Conv	A	10.45+4.65	8.98	5.92+3.90	4.53	3.70+4.04	1.90	26.50+11.91	25.50	5.67+4.43†	4.08
Trans	A1	10.48+3.82	9.73	5.7+2.49	4.69	3.78+4.36	1.79	26.81+10.46	25.53	5.62+3.95	4.58
Conv	A1	11.23+4.82†	9.42	6.65+4.91	5.11	3.74+4.04	1.76	28.14+11.35†	26.73	6.37+5.02†	4.72
Trans	A2	11.57+5.1	10.1	6.45+3.76	5.39	4.39+4.14	2.46	26.68+11.93	26.79	8.74+9.39	5.54
Conv	A2	11.65+5.45	9.94	6.91+4.58†	5.86	5.03+5.59	2.43	26.84+12.4	27.01	7.8+5.79	5.6
Trans	A+B	10.2+3.15	9.81	5.31+1.67	5.05	2.51+1.73	1.95	28.32+11.27	27.66	4.67+2.21	4.4
Conv	A+B	9.79+3.32†	9.32	4.83+1.54†	4.5	2.21+1.33†	1.79	27.85+11.9	27.0	4.26+2.27	3.74
Trans	A1+B	10.24+3.0	9.81	5.47+1.68	5.24	2.86+3.18	1.97	28.14+10.25	27.07	4.48+2.02	3.94
Conv	A1+B	9.87+3.19†	9.21	5.05+1.56†	4.75	2.19+1.38†	1.73	28.03+11.64	27.0	4.2+2.18	3.82
Trans	A2+B	10.4+3.22	10.24	5.31+1.49	4.86	2.48+1.4	2.15	28.11+11.19	28.36	5.69+3.0	5.15
Conv	A2+B	10.24+3.32	9.82	4.83+1.47†	4.53	2.32+1.23†	1.91	28.0+12.23	28.26	5.8+2.87	5.44
Trans⋆	A	10.42+3.8	9.41	5.51+2.37	4.8	2.85+2.67	1.79	27.61+11.63	26.84	5.72+4.24	4.3
Conv⋆	A	9.9+3.89†	8.89	5.12+3.12†	4.42	2.62+2.45†	1.76	26.84+11.82†	25.35	5.01+3.69†	3.83
Trans⋆	A1	10.51+3.89	9.4	5.55+2.46	4.63	3.35+3.98	1.77	27.32+10.67	26.44	5.8+3.89	4.54
Conv⋆	A1	10.29+3.95†	9.0	5.41+3.9†	4.5	2.85+3.51†	1.71	27.79+11.37	26.94	5.13+3.46†	4.18
Trans⋆	A2	10.69+4.22	9.59	5.78+3.0	4.87	3.41+3.03	2.21	27.28+12.06	26.19	6.29+4.9	5.01
Conv⋆	A2	10.87+5.27	9.46	5.7+3.66	4.6	3.64+3.74	2.23	26.7+12.46	26.8	7.45+8.73†	4.49

Table 4: Experiment 1: Replacing Swin-Transformer block with convolution block after halving the parameters. The Dice mean and standard deviation on EMC Test set(clinic C). ⋆ denotes the pretraining on clinic B and finetuning on clinic A. † represents statistical significant difference($p<0.05$ between the annotated method and Trans with the same training set

A.2. Experiments 2: Replacing the Self-Attention with Pooling

The test experiment results for replacing the Self-Attention with pooling are shown in table 5 and table 6. The experiment results for comparing average-pooling and max-pooling are shown in table 7 and table 8.

A.3. Experiments 3: Evaluating Positional Encoding

The test experiment results for evaluating positional encoding are shown in table 9 and 10. The experiment results for comparing learned absolute position embedding and sinsusoid position embedding are shown in table 11 and 12.

		Mean		Prostate		Bladder		Rectum		SeminalVesicle	
Network	Training Set	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median
Trans	A	81.82+5.07	83.15	83.83+4.43	85.44	95.65+3.24	96.66	74.31+9.66	74.48	73.48+15.59	77.45
Pool	A	82.38+4.66	84.02	84.61+4.3†	85.38	95.74+2.95	96.54	74.72+7.86	74.92	74.44+14.4	79.12
Trans	A1	81.31+5.12	82.31	83.44+4.25	84.11	95.69+3.4	96.75	72.8+10.81	72.54	73.3+14.29	76.77
Pool	A1	82.17+4.99†	83.56	83.9+4.3	84.42	96.04+2.85†	96.83	72.96+11.34	72.9	75.78+13.81†	78.38
Trans	A2	78.77+6.61	80.32	80.21+7.97	83.04	94.23+3.92	95.59	73.94+8.75	74.44	66.69+18.93	74.68
Pool	A2	80.05+6.74†	82.8	81.73+6.71†	83.6	94.58+3.71†	95.87	75.15+7.6†	75.62	68.74+18.28†	73.71
Trans	A+B	82.81+3.97	83.28	84.56+5.33	86.08	95.58+2.56	96.26	73.29+7.5	73.26	77.81+9.96	79.05
Pool	A+B	82.17+4.71	83.57	84.9+5.17	86.64	95.54+2.88	96.41	73.34+7.6	73.34	74.92+13.58	77.25†
Trans	A1+B	82.82+3.94	83.41	84.53+5.69	86.24	96.08+2.41	96.62	73.24+7.61	72.68	77.43+10.27	78.48
Pool	A1+B	82.46+4.18	83.31	84.29+5.61	86.13	96.15+2.25	96.75	73.23+7.49	72.72	76.15+10.94	76.82†
Trans	A2+B	81.95+4.19	82.91	84.88+5.08	86.53	95.29+2.3	95.83	72.8+8.02	72.95	74.84+10.83	75.14
Pool	A2+B	81.66+4.36	82.91	85.18+5.05	86.93	94.95+2.76†	95.68	73.47+7.48	73.37	73.05+10.8†	74.12
Trans⋆	A	82.89+4.14	84.07	84.92+4.31	86.47	95.7+3.25	96.55	74.87+7.18	74.29	76.08+12.96	78.68
Pool⋆	A	81.81+5.29†	83.49	84.28+4.11†	84.98	95.54+3.36	96.7	73.45+10.32	74.0	73.97+15.32†	78.74
Trans⋆	A1	82.18+4.69	83.44	84.27+4.1	85.36	95.93+3.37	97.1	74.13+7.66	73.25	74.41+14.85	78.59
Pool⋆	A1	81.97+5.31	83.56	83.95+3.88	84.65	95.74+3.25†	96.62	72.85+10.88†	72.43	75.32+14.87	79.06
Trans⋆	A2	80.9+5.39	82.78	83.02+5.6	85.09	94.69+3.69	95.78	74.45+7.53	74.83	71.45+16.22	75.86
Pool⋆	A2	79.76+6.95†	81.14	82.19+6.4†	83.97	94.51+3.76	95.88	73.55+10.87	74.97	68.78+17.46†	73.28

Table 5: Experiments 2: Replacing the Self-Attention with Pooling. The Dice mean and standard deviation on EMC Test set(clinic C). $\star$ denotes the pretraining on clinic B and finetuning on clinic A .† represents statistical significant difference($p<0.05$ between the annotated method and Trans with the same training set.

		Mean		Prostate		Bladder		Rectum		SeminalVesicle	
Network	Training Set	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median
Trans	A	10.31+3.77	9.46	5.42+2.38	4.5	2.98+2.57	1.89	26.96+11.64	26.61	5.88+4.39	4.49
Pool	A	10.16+3.79	9.59	5.26+2.39†	4.76	2.78+2.26	1.83	26.91+11.7	26.28	5.68+4.28	4.57
Trans	A1	10.37+3.59	9.5	5.5+2.25	4.87	3.2+2.76	1.89	27.17+10.63	25.91	5.61+3.84	4.64
Pool	A1	10.03+3.5†	9.3	5.42+2.28	4.5	2.92+2.47	1.79	26.71+11.05†	25.51	5.06+4.11†	3.89
Trans	A2	11.18+4.56	9.99	6.58+3.54	5.73	4.31+3.94	2.59	26.93+11.96	26.71	6.89+4.94	4.92
Pool	A2	10.84+4.48†	9.84	6.19+3.53†	5.17	3.74+3.11†	2.26	27.16+11.75	27.01	6.26+4.66†	4.92
Trans	A+B	9.95+3.13	9.54	5.11+1.56	4.69	2.42+1.65	1.83	27.82+11.24	27.0	4.45+2.21	4.09
Pool	A+B	10.35+3.52	9.94	5.15+2.16	4.5	2.61+2.14	1.95	28.49+11.98	28.5	5.15+3.15†	4.48
Trans	A1+B	10.11+3.1	9.53	5.2+1.59	4.88	2.3+1.76	1.79	28.32+11.15	27.27	4.61+2.44	4.11
Pool	A1+B	10.31+3.12	9.66	5.36+1.63	5.05	2.29+1.55	1.83	28.82+11.46	27.88	4.77+2.17	4.78
Trans	A2+B	10.35+3.31	9.89	5.09+1.5	4.58	2.44+1.41	1.99	28.55+11.98	28.26	5.33+2.64	4.64
Pool	A2+B	10.77+4.53	10.05	4.95+1.91	4.53	2.69+1.92	2.19	28.42+12.11	28.59	7.01+9.72	5.6
Trans⋆	A	10.03+3.54	9.2	5.17+2.39	4.4	2.74+2.58	1.76	27.33+11.69	26.74	4.89+3.09	3.94
Pool⋆	A	10.34+3.84†	9.89	5.25+2.55	4.58	3.07+2.72†	1.89	27.6+11.93	28.5	5.43+3.24†	4.43
Trans⋆	A1	10.41+3.84	9.38	5.45+2.53	4.5	3.29+3.87	1.77	27.44+10.91	26.66	5.49+3.61	4.66
Pool⋆	A1	10.31+3.55	9.85	5.57+2.76	4.5	3.3+3.02	1.79	27.41+10.62	27.32	4.95+2.92	4.46
Trans⋆	A2	10.68+4.16	9.75	5.74+3.16	4.67	3.46+3.13	2.23	27.24+11.85†	26.87	6.3+4.96	4.82
Pool⋆	A2	10.98+4.59†	10.27	6.04+3.44†	5.18	3.85+3.37†	2.39	27.6+12.3†	28.12	6.43+4.56	5.07

Table 6: Experiments 2: Replacing the Self-Attention with Pooling. The HD95 mean and standard deviation on EMC Test set(clinic C). $\star$ denotes the pretraining on clinic B and finetuning on clinic A .† represents statistical significant difference($p<0.05$ between the annotated method and Trans with the same training set.

		Mean		Prostate		Bladder		Rectum		SeminalVesicle	
Network	Training Set	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median
AvgPooling	A	80.92+6.37	83.03	83.13+5.55	84.51	95.47+3.04	96.56	74.68+8.31	74.8	70.39+20.0	78.88
MaxPooling	A	82.38+4.66†	84.02	84.61+4.3†	85.38	95.74+2.95†	96.54	74.72+7.86	74.92	74.44+14.4	79.12
AvgPooling	A1	80.26+6.33	82.96	82.22+6.03	84.33	95.71+3.38	96.86	72.11+10.5	72.52	70.98+16.89	77.27
MaxPooling	A1	82.17+4.99†	83.56	83.9+4.3†	84.42	96.04+2.85†	96.83	72.96+11.34†	72.9	75.78+13.81†	78.38
AvgPooling	A2	78.1+8.11	80.58	80.11+8.44	82.96	94.25+3.8	95.81	74.32+8.6	74.98	63.71+22.98	76.09
MaxPooling	A2	80.05+6.74†	82.8	81.73+6.71†	83.6	94.58+3.71†	95.87	75.15+7.6†	75.62	68.74+18.28†	73.71

Table 7: Comparing average-pooling with max-pooling in small data scales. The Dice mean and standard deviation on EMC Test set(clinic C). † represents statistical significant difference($p<0.05$ between the annotated method and Trans with the same training set

		Mean		Prostate		Bladder		Rectum		SeminalVesicle	
Network	Training Set	$\mu+\sigma$	median	$\mu+\sigma$	median	$\mu+\sigma$	median	$\mu+\sigma$	median	$\mu+\sigma$	median
AvgPooling	A	10.66+4.52	9.46	5.52+2.89	4.55	3.18+2.56	1.9	26.94+11.97	27.0	7.0+6.18	4.34
MaxPooling	A	10.16+3.79	9.59	5.26+2.39	4.76	2.78+2.26†	1.83	26.91+11.7	26.28	5.68+4.28	4.57
AvgPooling	A1	10.77+4.18	9.79	6.0+3.15	5.02	3.34+2.96	1.79	27.27+10.42	26.92	6.48+4.87	4.79
MaxPooling	A1	10.03+3.5†	9.3	5.42+2.28†	4.5	2.92+2.47†	1.79	26.71+11.05†	25.51	5.06+4.11†	3.89
AvgPooling	A2	11.66+5.39	10.17	6.46+3.66	5.67	4.18+3.68	2.46	26.77+12.07	27.0	9.24+9.95	4.75
MaxPooling	A2	10.84+4.48†	9.84	6.19+3.53	5.17	3.74+3.11†	2.26	27.16+11.75	27.01	6.26+4.66†	4.92

Table 8: Comparing average-pooling with max-pooling in small data scales. The HD95 mean and standard deviation on EMC Test set(clinic C). † represents statistical significant difference($p<0.05$ between the annotated method and Trans with the same training set

		Mean		Prostate		Bladder		Rectum		SeminalVesicle	
Network	Training Set	$\mu+\sigma$	median	$\mu+\sigma$	median	$\mu+\sigma$	median	$\mu+\sigma$	median	$\mu+\sigma$	median
Trans w/o	A	81.98+4.61	82.59	84.01+4.29	85.46	95.7+3.06	96.69	74.37+8.99	74.51	73.84+14.21	77.68
TransR	A	81.82+5.07	83.15	83.83+4.43	85.44	95.65+3.24	96.66	74.31+9.66	74.48	73.48+15.59	77.45
TransA	A	81.4+5.28†‡	82.47	83.78+4.57	85.31	95.58+3.45	96.81	74.15+9.05	73.66	72.09+16.51†‡	76.87
Trans w/o	A1	81.21+5.32	82.32	83.37+4.56	84.24	95.63+3.64	96.81	72.84+10.62	72.54	72.99+15.04	76.46
TransR	A1	81.31+5.12	82.31	83.44+4.25	84.11	95.69+3.4	96.75	72.8+10.81	72.54	73.3+14.29	76.77
TransA	A1	81.39+5.08	82.42	83.67+4.2	84.26	95.75+3.46	96.99	72.97+10.62	72.87	73.18+14.06	75.76
Trans w/o	A2	79.16+6.52	80.72	80.34+8.2	83.43	94.29+4.05	95.71	74.13+8.01	75.07	67.88+17.97	73.65
TransR	A2	78.77+6.61†	80.32	80.21+7.97	83.04	94.23+3.92†	95.59	73.94+8.75	74.44	66.69+18.93	74.68
TransA	A2	78.62+6.94†	80.49	80.15+8.18	82.86	94.29+3.98	95.76	73.92+8.43	74.4	66.12+19.57†	73.13
Trans w/o	A+B	82.64+4.06	83.24	84.55+5.57	86.35	95.59+2.67	96.19	73.16+7.7	72.8	77.27+10.22	78.36
TransR	A+B	82.81+3.97†	83.28	84.56+5.33	86.08	95.58+2.56	96.26	73.29+7.5	73.26	77.81+9.96†	79.05
TransA	A+B	82.81+4.01	83.69	84.91+5.33‡	86.59	95.58+2.71	96.28	73.57+7.43†	73.36	77.19+10.32	78.17
Trans w/o	A1+B	82.62+4.06	82.92	84.01+5.98	85.55	95.99+2.73	96.63	73.1+7.45	72.46	77.4+9.98	78.75
TransR	A1+B	82.82+3.94†	83.41	84.53+5.69†	86.24	96.08+2.41†	96.62	73.24+7.61	72.68	77.43+10.27	78.48
TransA	A1+B	82.11+4.0†‡	82.47	84.1+6.06‡	86.16	95.78+2.3†‡	96.35	72.17+7.71†‡	71.55	76.41+10.04†‡	77.42
Trans w/o	A2+B	82.15+3.96	82.62	84.86+4.88	85.51	95.57+2.17	96.0	73.19+7.55	73.08	74.98+10.03	74.89
TransR	A2+B	81.95+4.19	82.91	84.88+5.08	86.53	95.29+2.3†	95.83	72.8+8.02†	72.95	74.84+10.83	75.14
TransA	A2+B	81.92+3.98†‡	82.37	84.49+5.41†‡	85.51	95.23+2.28†‡	95.76	72.89+7.43†	72.62	75.07+9.85	75.77
Trans w/o ⋆	A	82.57+4.31	83.6	84.57+4.45	86.13	95.88+2.83	96.58	74.53+7.35	73.89	75.28+13.76	78.33
TransR⋆	A	82.89+4.14†	84.07	84.92+4.31†	86.47	95.7+3.25†	96.55	74.87+7.18†	74.29	76.08+12.96†	78.68
TransA⋆	A	82.69+4.62	83.79	84.58+4.54‡	86.6	95.73+3.22†	96.52	74.92+7.27†	74.4	75.54+14.63	79.66
Trans w/o ⋆	A1	82.32+4.54	83.76	84.09+4.49	85.91	96.03+3.27	97.18	73.95+7.69	72.77	75.22+13.97	78.46
TransR⋆	A1	82.18+4.69	83.44	84.27+4.1	85.36	95.93+3.37	97.1	74.13+7.66	73.25	74.41+14.85	78.59
TransA⋆	A1	82.43+4.63	83.6	84.24+4.52	85.3	95.98+3.44	97.12	74.34+7.48†	73.68	75.15+14.16	78.7
Trans w/o ⋆	A2	81.11+5.08	82.59	83.25+5.23	85.56	94.7+3.47	95.76	74.53+7.41	74.24	71.96+15.42	76.39
TransR⋆	A2	80.9+5.39	82.78	83.02+5.6	85.09	94.69+3.69	95.78	74.45+7.53	74.83	71.45+16.22	75.86
TransA⋆	A2	81.12+5.22	82.61	83.54+5.17†‡	85.32	94.79+3.59†‡	95.82	74.62+7.48	74.3	71.54+16.08	75.52

Table 9: Experiments 3: Evaluating Positional Encoding. The Dice mean and standard deviation on EMC Test set(clinic C). † represents statistical significant difference($p<0.05$ between the annotated method and Transformer without Positional encoding with the same training set. ‡ represents statistical significant difference($p<0.05$ between the annotated method and Transformer with relative positional bias(Trans) with the same training set.

A.4. Side Experiments

The test experiment results for the comparison between models of different size(75m,150m,300m) are shown in table 13 and table 14. The experiment results for the comparison between models with/without Layer Normalization are shown in table 15 and table 16.

B. Validation Experiment results

B.1. Experiment1: Replacing Swin-Transformer Block with Convolution

The validation experiment results for replacing Swin-Transformer Block with Convolution are shown in table 17 and table 18. The results after halving the parameters size are shown in table 19 and table 20.

B.2. Experiment 2: Replacing the Self-Attention with Pooling

The validation experiment results for replacing the Self-Attention with pooling are shown in table 21 and table 22. The experiment results for comparing average-pooling and max-pooling are shown in table 23 and table 24.

		Mean		Prostate		Bladder		Rectum		SeminalVesicle	
Network	Training Set	$\mu+\sigma$	median	$\mu+\sigma$	median	$\mu+\sigma$	median	$\mu+\sigma$	median	$\mu+\sigma$	median
Trans w/o	A	10.48+3.92	9.66	5.37+2.41	4.7	3.42+4.62	1.79	27.41+11.76	27.0	5.72+3.89	4.36
TransR	A	10.31+3.77†	9.46	5.42+2.38	4.5	2.98+2.57	1.89	26.96+11.64†	26.61	5.88+4.39	4.49
TransA	A	10.48+4.06	9.17	5.48+2.6	4.62	3.44+4.18‡	1.79	26.97+11.47†	25.5	6.03+4.48	4.67
Trans w/o	A1	10.48+3.88	9.32	5.74+2.74	4.78	3.67+3.97	1.91	27.03+10.42	25.52	5.5+3.83	4.6
TransR	A1	10.37+3.59	9.5	5.5+2.25	4.87	3.2+2.76†	1.89	27.17+10.63	25.91	5.61+3.84	4.64
TransA	A1	10.35+3.79†	9.42	5.56+2.54	4.74	3.3+3.1	1.85	27.03+10.74	25.52	5.52+3.68	4.62
Trans w/o	A2	11.15+4.48	10.06	6.59+3.86	5.61	4.17+3.98	2.46	27.22+11.85†	27.04	6.61+4.68	5.35
TransR	A2	11.18+4.56	9.99	6.58+3.54	5.73	4.31+3.94	2.59	26.93+11.96	26.71	6.89+4.94	4.92
TransA	A2	11.33+4.66†	9.95	6.58+3.9	5.77	4.3+4.13	2.46	27.32+12.07‡	27.02	7.13+5.28†	5.19
Trans w/o	A+B	10.11+3.34	9.55	5.14+1.73	4.57	2.43+1.88	1.83	28.29+11.6	27.03	4.58+2.29	4.2
TransR	A+B	9.95+3.13†	9.54	5.11+1.56	4.69	2.42+1.65	1.83	27.82+11.24†	27.0	4.45+2.21†	4.09
TransA	A+B	9.95+3.18†	9.34	5.0+1.65†	4.6	2.42+1.85	1.8	27.81+11.16†	27.0	4.58+2.28	4.19
Trans w/o	A1+B	10.07+3.17	9.4	5.37+1.83	5.01	2.4+2.1	1.79	27.94+10.68	27.0	4.56+2.17	4.38
TransR	A1+B	10.11+3.1	9.53	5.2+1.59	4.88	2.3+1.76	1.79	28.32+11.15	27.27	4.61+2.44	4.11
TransA	A1+B	10.08+3.05	9.62	5.18+1.62	4.8	2.46+1.63†‡	1.91	28.02+10.77	27.07	4.67+2.18	4.31
Trans w/o	A2+B	10.33+3.31	9.82	5.11+1.35	4.92	2.35+1.25	1.95	28.48+11.8	28.39	5.37+2.75	5.03
TransR	A2+B	10.35+3.31	9.89	5.09+1.5	4.58	2.44+1.41†	1.99	28.55+11.98	28.26	5.33+2.64	4.64
TransA	A2+B	10.27+3.14	9.92	5.18+1.54	4.85	2.47+1.39	2.07	28.21+11.23	27.89	5.22+2.51	4.51
Trans w/o ⋆	A	10.16+3.49	9.41	5.25+1.98	4.73	2.61+2.11	1.79	27.44+11.64	26.49	5.33+3.5	4.54
TransR⋆	A	10.03+3.54†	9.2	5.17+2.39†	4.4	2.74+2.58	1.76	27.33+11.69	26.74	4.89+3.09†	3.94
TransA⋆	A	10.19+3.66‡	9.32	5.27+2.32	4.5	2.72+2.58	1.76	27.57+11.79‡	25.73	5.2+3.6	4.19
Trans w/o ⋆	A1	10.36+3.75	9.2	5.46+2.26	4.55	3.21+3.85	1.77	27.5+10.75	27.0	5.29+3.5	4.36
TransR⋆	A1	10.41+3.84	9.38	5.45+2.53	4.5	3.29+3.87†	1.77	27.44+10.91†	26.66	5.49+3.61	4.66
TransA⋆	A1	10.24+3.75	9.27	5.41+2.39	4.53	3.22+3.61	1.79	27.06+10.52†‡	25.96	5.28+3.57	4.6
Trans w/o ⋆	A2	10.61+4.09	9.81	5.64+2.95	4.69	3.36+2.85	2.23	27.31+11.94	27.0	6.14+4.58	4.72
TransR⋆	A2	10.68+4.16	9.75	5.74+3.16	4.67	3.46+3.13†	2.23	27.24+11.85	26.87	6.3+4.96	4.82
TransA⋆	A2	10.73+4.3	9.69	5.54+3.03‡	4.57	3.28+2.89‡	2.22	27.39+11.85‡	26.88	6.69+7.48	4.89

Table 10: Experiments 3: Evaluating Positional Encoding. The HD95 mean and standard deviation on EMC Test set(clinic C). † represents statistical significant difference(p<0.05 between the annotated method and Transformer without Positional encoding with the same training set. ‡ represents statistical significant difference(p<0.05 between the annotated method and Transformer with relative positional bias(Trans) with the same training set.

		Mean		Prostate		Bladder		Rectum		SeminalVesicle	
Network	Training Set	$\mu+\sigma$	median	$\mu+\sigma$	median	$\mu+\sigma$	median	$\mu+\sigma$	median	$\mu+\sigma$	median
TransA	A	81.4+5.28	82.47	83.78+4.57	85.31	95.58+3.45	96.81	74.15+9.05	73.66	72.09+16.51	76.87
TransS	A	81.59+4.88	82.49	83.57+4.91	85.54	95.58+3.31	96.7	74.16+8.26	73.56	73.02+15.04	76.93
TransA	A1	81.39+5.08	82.42	83.67+4.2	84.26	95.75+3.46	96.99	72.97+10.62	72.87	73.18+14.06	75.76
TransS	A1	81.15+5.32	82.45	83.44+4.25	84.59	95.76+3.27	96.82	72.7+10.41†	72.36	72.71+15.47	76.65
TransA	A2	78.62+6.94	80.49	80.15+8.18	82.86	94.29+3.98	95.76	73.92+8.43	74.4	66.12+19.57	73.13
TransS	A2	79.1+6.76†	80.63	79.92+8.23	82.67	94.35+3.95†	95.86	74.02+8.46	74.96	68.12+18.56†	74.54
TransA	A+B	82.81+4.01	83.69	84.91+5.33	86.59	95.58+2.71	96.28	73.57+7.43	73.36	77.19+10.32	78.17
TransS	A+B	82.58+3.97†	83.33	84.73+5.46	86.01	95.48+2.62†	96.13	72.91+7.43†	72.73	77.21+9.94	77.56
TransA	A1+B	82.11+4.0	82.47	84.1+6.06	86.16	95.78+2.3	96.35	72.17+7.71	71.55	76.41+10.04	77.42
TransS	A1+B	82.58+4.01†	83.08	84.45+6.06	86.11	95.77+2.48	96.37	72.72+7.48†	71.65	77.37+9.9†	77.63
TransA	A2+B	81.92+3.98	82.37	84.49+5.41	85.51	95.23+2.28	95.76	72.89+7.43	72.62	75.07+9.85	75.77
TransS	A2+B	81.63+4.26	81.82	84.88+5.36†	86.83	95.13+2.26	95.74	72.89+7.55†	72.75	73.61+12.03	75.47
TransA⋆	A	82.69+4.62	83.79	84.58+4.54	86.6	95.73+3.22	96.52	74.92+7.27	74.4	75.54+14.63	79.66
TransS⋆	A	82.72+4.2	83.8	84.93+4.19†	86.3	95.76+3.14	96.63	74.6+7.34†	74.23	75.6+12.84	78.37
TransA⋆	A1	82.43+4.63	83.6	84.24+4.52	85.3	95.98+3.44	97.12	74.34+7.48	73.68	75.15+14.16	78.7
TransS⋆	A1	82.28+4.52	83.35	84.26+4.42	85.51	95.92+3.43	97.04	73.94+7.55	73.72	75.0+13.4	78.49
TransA⋆	A2	81.12+5.22	82.61	83.54+5.17	85.32	94.79+3.59	95.82	74.62+7.48	74.3	71.54+16.08	75.52
TransS⋆	A2	81.09+5.31	82.53	83.5+5.06	85.0	94.84+3.57†	95.9	74.63+7.5†	74.73	71.41+16.32	75.79

Table 11: Comparison between learned and unlearned sinusoid absolute position embedding. The Dice mean and standard deviation on EMC Test set(clinic C). ⋆ denotes the pretraining on clinic B and finetuning on clinic A .† represents statistical significant difference(p<0.05 between the annotated method and Trans with the same training set.

B.3. Experiment 3: Evaluating Positional Encoding

The validation experiment results for evaluating positional encoding are shown in table **??** and 26. The experiment results for comparing learned absolute position embedding and sinsusoid position embedding are shown in table 27 and 28.

		Mean		Prostate		Bladder		Rectum		SeminalVesicle	
Network	Training Set	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median
TransA	A	10.48+4.06	9.17	5.48+2.6	4.62	3.44+4.18	1.79	26.97+11.47	25.5	6.03+4.48	4.67
TransS	A	10.39+3.89	9.33	5.5+2.66	4.58	3.19+2.99	1.95	27.12+11.6	25.95	5.74+3.96†	4.27
TransA	A1	10.35+3.79	9.42	5.56+2.54	4.74	3.3+3.1	1.85	27.03+10.74	25.52	5.52+3.68	4.62
TransS	A1	10.48+3.73	9.44	5.65+2.48	4.72	3.3+2.97	1.85	27.46+10.99†	25.51	5.49+3.93	4.67
TransA	A2	11.33+4.66	9.95	6.58+3.9	5.77	4.3+4.13	2.46	27.32+12.07	27.02	7.13+5.28	5.19
TransS	A2	11.21+4.56	9.99	6.7+3.96	5.72	4.43+4.39†	2.38	27.13+12.13	27.09	6.59+4.98†	5.01
TransA	A+B	9.95+3.18	9.34	5.0+1.65	4.6	2.42+1.85	1.8	27.81+11.16	27.0	4.58+2.28	4.19
TransS	A+B	10.01+3.25	9.68	5.02+1.71	4.67	2.48+1.81	1.94	28.02+11.39	27.49	4.54+2.17	4.09
TransA	A1+B	10.08+3.05	9.62	5.18+1.62	4.8	2.46+1.63	1.91	28.02+10.77	27.07	4.67+2.18	4.31
TransS	A1+B	9.98+3.0	9.57	5.09+1.69	4.55	2.47+1.81	1.95	27.93+10.67	27.19	4.43+2.16†	4.23
TransA	A2+B	10.27+3.14	9.92	5.18+1.54	4.85	2.47+1.39	2.07	28.21+11.23	27.89	5.22+2.51	4.51
TransS	A2+B	10.24+3.33	9.77	4.97+1.57†	4.57	2.51+1.37†	2.18	27.85+11.65	27.0	5.63+3.11	4.7
TransA⋆	A	10.19+3.66	9.32	5.27+2.32	4.5	2.72+2.58	1.76	27.57+11.79	25.73	5.2+3.6	4.19
TransS⋆	A	10.16+3.6	9.25	5.11+2.33†	4.5	2.79+2.5	1.79	27.6+11.75	26.96	5.16+3.19	4.23
TransA⋆	A1	10.24+3.75	9.27	5.41+2.39	4.53	3.22+3.61	1.79	27.06+10.52	25.96	5.28+3.57	4.6
TransS⋆	A1	10.37+3.85	9.24	5.45+2.72	4.5	3.42+4.07	1.79	27.49+10.83†	26.68	5.13+3.02	4.51
TransA⋆	A2	10.73+4.3	9.69	5.54+3.03	4.57	3.28+2.89	2.22	27.39+11.85	26.88	6.69+7.48	4.89
TransS⋆	A2	10.6+4.09	9.7	5.59+2.94	4.89	3.3+2.92	2.21	27.21+11.97†	26.59	6.31+4.71	4.81

Table 12: Comparison between learned and unlearned sinusoid absolute position embedding. The HD95 mean and standard deviation on EMC Test set(clinic C). $\star$ denotes the pretraining on clinic B and finetuning on clinic A .† represents statistical significant difference($p<0.05$ between the annotated method and Trans with the same training set.

		Mean		Prostate		Bladder		Rectum		SeminalVesicle	
Network	Training Set	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median
Trans	A2	78.77+6.61	80.32	80.21+7.97	83.04	94.23+3.92	95.59	73.94+8.75	74.44	66.69+18.93	74.68
Conv	A2	77.9+7.8	80.15	80.84+8.94	84.07	94.01+4.55	95.77	75.02+7.7	74.37	61.75+20.71	69.37
Trans-L	A2	79.52+6.35†‡	81.11	81.69+7.11†	84.27	94.65+3.91†‡	95.99	74.83+8.23†	75.48	66.91+18.31‡	72.83
Trans-S	A2	78.06+7.47†	80.05	80.15+8.53‡	83.36	94.21+4.08	95.67	73.65+9.27‡	73.69	64.21+21.09†‡	72.49
Trans	A2+B	81.95+4.19	82.91	84.88+5.08	86.53	95.29+2.3	95.83	72.8+8.02	72.95	74.84+10.83	75.14
Conv	A2+B	82.68+4.07	83.34	86.28+4.47	87.54	95.64+2.21	95.98	73.94+7.49	73.75	74.84+10.86	76.26
Trans-L	A2+B	82.31+3.9‡	82.89	85.24+4.99‡	86.95	95.37+2.3‡	95.91	73.98+7.68	74.03	74.66+9.77	75.54
Trans-S	A2+B	80.98+4.57†‡	81.73	84.02+5.5†‡	84.9	95.2+2.35†‡	95.68	72.19+8.09‡	72.43	72.5+12.38†‡	74.91
Trans	A1	81.31+5.12	82.31	83.44+4.25	84.11	95.69+3.4	96.75	72.8+10.81	72.54	73.3+14.29	76.77
Conv	A1	80.51+6.55	82.08	81.46+9.95	84.46	95.8+3.57	96.99	73.81+7.63	73.48	70.97+17.63	75.56
Trans-L	A1	81.11+5.42	82.87	83.15+4.96	84.42	95.76+3.6†	96.94	73.01+10.75	72.74	72.53+15.54‡	77.49
Trans-S	A1	81.06+5.38	82.59	83.17+4.39	84.14	95.68+3.37‡	96.93	72.8+10.94	72.57	72.58+15.0	77.05
Trans	A1+B	82.82+3.94	83.41	84.53+5.69	86.24	96.08+2.41	96.62	73.24+7.61	72.68	77.43+10.27	78.48
Conv	A1+B	83.69+3.95	84.36	85.6+5.11	87.11	96.58+1.72	96.96	74.06+7.59	73.47	78.52+10.45	79.17
Trans-L	A1+B	83.41+3.94†	83.81	85.15+5.26†‡	86.83	96.34+2.04†‡	96.89	73.48+7.74‡	73.12	78.66+10.21†	79.44
Trans-S	A1+B	81.7+4.1†‡	82.15	83.43+6.02†‡	85.09	95.51+2.48†‡	96.01	71.85+7.52†‡	71.07	76.02+10.21†‡	77.88
Trans	A	81.82+5.07	83.15	83.83+4.43	85.44	95.65+3.24	96.66	74.31+9.66	74.48	73.48+15.59	77.45
Conv	A	81.73+5.77	83.53	83.57+6.29	85.81	95.45+3.79	96.77	75.75+7.45	76.09	72.14+17.07	78.27
Trans-L	A	81.75+4.85	83.12	83.47+4.93†	84.88	95.77+3.3	96.79	74.49+8.65‡	74.52	73.27+15.6	77.79
Trans-S	A	81.45+5.15†	82.68	83.59+4.76	85.24	95.57+3.18†‡	96.6	73.56+9.57†‡	73.24	73.08+15.71	78.35
Trans	A+B	82.81+3.97	83.28	84.56+5.33	86.08	95.58+2.56	96.26	73.29+7.5	73.26	77.81+9.96	79.05
Conv	A+B	83.95+3.88	84.57	86.59+4.57	87.72	96.13+1.85	96.55	74.55+7.52	74.39	78.55+10.48	80.39
Trans-L	A+B	83.42+3.92†‡	84.09	85.3+4.91†‡	86.18	95.82+2.4†‡	96.51	74.03+7.44†	73.93	78.51+9.88	80.05
Trans-S	A+B	82.01+4.02†‡	82.77	83.98+5.68†‡	84.6	95.41+2.54†‡	96.07	72.17+7.79†‡	71.94	76.47+9.57†‡	77.56

Table 13: Comparison between models of different size. The Dice mean and standard deviation on EMC Test set(clinic C). † represents statistical significant difference($p<0.05$ between the annotated method and Transformer with the same training set. ‡ represents statistical significant difference($p<0.05$ between the annotated method and the convolution with the same training set.

B.4. Side Experiments

The validation experiment results for the comparison between models of different size(75m,150m,300m) are shown in table 29 and table 29. The experiment results for the comparison between models with/without Layer Normalization are shown in table 31 and table 20.

		Mean		Prostate		Bladder		Rectum		SeminalVesicle	
Network	Training Set	$\mu+\sigma$	median	$\mu+\sigma$	median	$\mu+\sigma$	median	$\mu+\sigma$	median	$\mu+\sigma$	median
Trans	A2	11.18+4.56	9.99	6.58+3.54	5.73	4.31+3.94	2.59	26.93+11.96	26.71	6.89+4.94	4.92
Conv	A2	11.51+5.27	10.09	6.6+4.26	5.2	4.66+4.96	2.41	26.7+12.41	27.0	8.08+5.86	5.97
Trans-L	A2	11.05+4.47	10.21	6.14+3.55†‡	5.06	3.86+3.73†‡	2.2	27.05+12.11	26.77	7.14+5.1‡	5.38
Trans-S	A2	11.57+5.1	10.1	6.45+3.76	5.39	4.39+4.14	2.46	26.68+11.93†	26.79	8.74+9.39	5.54
Trans	A2+B	10.35+3.31	9.89	5.09+1.5	4.58	2.44+1.41	1.99	28.55+11.98	28.26	5.33+2.64	4.64
Conv	A2+B	10.21+3.4	9.71	4.85+1.56	4.51	2.33+1.42	1.88	27.9+12.12	27.01	5.75+2.93	5.69
Trans-L	A2+B	10.36+3.37‡	9.9	5.11+1.55‡	4.78	2.39+1.49	1.95	28.09+12.04	27.36	5.85+3.09	5.16
Trans-S	A2+B	10.4+3.22‡	10.24	5.31+1.49†‡	4.86	2.48+1.4‡	2.15	28.11+11.19	28.36	5.69+3.0	5.15
Trans	A1	10.37+3.59	9.5	5.5+2.25	4.87	3.2+2.76	1.89	27.17+10.63	25.91	5.61+3.84	4.64
Conv	A1	10.99+4.52	9.35	6.45+4.56	5.06	3.35+3.35	1.76	28.06+11.32	27.0	6.11+4.83	4.67
Trans-L	A1	10.58+3.94†‡	9.46	5.74+2.66	4.67	3.31+3.12	1.84	27.47+11.33	25.89	5.81+4.19	4.63
Trans-S	A1	10.48+3.82‡	9.73	5.7+2.49†	4.69	3.78+4.36	1.79	26.81+10.46	25.53	5.62+3.95	4.58
Trans	A1+B	10.11+3.1	9.53	5.2+1.59	4.88	2.3+1.76	1.79	28.32+11.15	27.27	4.61+2.44	4.11
Conv	A1+B	9.82+3.15	9.32	4.94+1.4	4.71	2.08+1.0	1.72	28.05+11.65	27.0	4.21+2.14	3.9
Trans-L	A1+B	9.86+2.94†	9.4	4.96+1.45†	4.64	2.16+1.42†	1.72	27.92+10.72	26.87	4.41+2.41	4.01
Trans-S	A1+B	10.24+3.0‡	9.81	5.47+1.68†‡	5.24	2.86+3.18†‡	1.97	28.14+10.25	27.07	4.48+2.02	3.94
Trans	A	10.31+3.77	9.46	5.42+2.38	4.5	2.98+2.57	1.89	26.96+11.64	26.61	5.88+4.39	4.49
Conv	A	10.35+4.51	9.07	5.84+3.8	4.54	3.7+4.59	1.81	26.42+11.79	25.32	5.45+4.06	3.71
Trans-L	A	10.34+3.81	9.37	5.58+2.59†	4.84	2.94+2.6	1.79	27.21+11.8‡	26.47	5.64+4.01	4.54
Trans-S	A	10.41+3.84	9.76	5.47+2.48	4.76	3.09+2.63	1.95	27.43+11.68†‡	26.86	5.66+4.02	3.91
Trans	A+B	9.95+3.13	9.54	5.11+1.56	4.69	2.42+1.65	1.83	27.82+11.24	27.0	4.45+2.21	4.09
Conv	A+B	9.62+3.33	9.12	4.65+1.42	4.5	2.14+1.09	1.75	27.39+12.02	25.96	4.31+2.32	3.92
Trans-L	A+B	9.82+3.19†‡	9.32	4.9+1.45†‡	4.51	2.34+1.62†‡	1.79	27.56+11.56†	26.94	4.46+2.29	3.67
Trans-S	A+B	10.2+3.15†‡	9.81	5.31+1.67†‡	5.05	2.51+1.73†‡	1.95	28.32+11.27†‡	27.66	4.67+2.21†	4.4

Table 14: Comparison between models of different size. The HD95 mean and standard deviation on EMC Test set(clinic C). † represents statistical significant difference($p<0.05$ between the annotated method and Transformer with the same training set. ‡ represents statistical significant difference($p<0.05$ between the annotated method and the convolution with the same training set.

		Mean		Prostate		Bladder		Rectum		SeminalVesicle	
Network	Training Set	$\mu+\sigma$	median	$\mu+\sigma$	median	$\mu+\sigma$	median	$\mu+\sigma$	median	$\mu+\sigma$	median
Trans	A	81.82+5.07	83.15	83.83+4.43	85.44	95.65+3.24	96.66	74.31+9.66	74.48	73.48+15.59	77.45
Trans w/o LN	A	80.78+5.48†	82.48	82.76+4.94†	84.57	95.03+3.56†	96.39	72.34+11.25†	73.14	73.0+15.9	77.03
Trans	A1	81.31+5.12	82.31	83.44+4.25	84.11	95.69+3.4	96.75	72.8+10.81	72.54	73.3+14.29	76.77
Trans w/o LN	A1	80.59+5.82	81.99	83.34+4.04	84.02	94.95+4.39†	96.46	71.88+11.61†	72.94	72.2+15.71	76.18
Trans	A2	78.77+6.61	80.32	80.21+7.97	83.04	94.23+3.92	95.59	73.94+8.75	74.44	66.69+18.93	74.68
Trans w/o LN	A2	77.61+7.55†	80.12	77.36+8.25†	80.03	93.69+3.69†	95.08	71.63+10.98†	71.85	67.77+21.23	74.61
Trans	A+B	82.81+3.97	83.28	84.56+5.33	86.08	95.58+2.56	96.26	73.29+7.5	73.26	77.81+9.96	79.05
Trans w/o LN	A+B	79.65+4.13†	81.02	80.62+6.42†	81.82	94.11+3.55†	95.29	70.0+9.44†	70.08	73.87+11.21†	75.92
Trans	A1+B	82.82+3.94	83.41	84.53+5.69	86.24	96.08+2.41	96.62	73.24+7.61	72.68	77.43+10.27	78.48
Trans w/o LN	A1+B	79.78+4.22†	80.82	80.91+6.19†	82.22	94.73+3.26†	95.75	70.05+8.39†	69.85	73.42+13.5†	76.42
Trans	A2+B	80.47+4.24	81.29	83.22+5.78	84.7	94.47+2.46	95.02	70.41+7.21	70.46	73.78+10.81	74.92
Trans w/o LN	A2+B	78.55+5.52†	80.43	80.03+6.77†	80.74	94.1+3.54	95.31	70.21+8.32	70.03	69.85+14.84†	72.2

Table 15: Comparison between the Transformer with/without Layer Normalization. The HD95 mean and standard deviation on EMC test set(clinic C). † represents statistical significant difference($p<0.05$ between the annotated method and Transformer with Layer Normalization in the same training set.

		Mean		Prostate		Bladder		Rectum		SeminalVesicle	
Network	Training Set	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median
Trans	A	10.31+3.77	9.46	5.42+2.38	4.5	2.98+2.57	1.89	26.96+11.64	26.61	5.88+4.39	4.49
Trans w/o LN	A	11.35+6.19†	9.35	5.81+2.6†	4.86	5.75+17.44†	2.37	27.55+11.71	25.78	6.28+8.31	4.01
Trans	A1	10.37+3.59	9.5	5.5+2.25	4.87	3.2+2.76	1.89	27.17+10.63	25.91	5.61+3.84	4.64
Trans w/o LN	A1	10.85+4.7	9.59	5.59+2.43	4.72	4.47+5.16†	2.82	27.33+11.24	25.5	6.02+7.77	4.35
Trans	A2	11.18+4.56	9.99	6.58+3.54	5.73	4.31+3.94	2.59	26.93+11.96	26.71	6.89+4.94	4.92
Trans w/o LN	A2	12.19+6.65	9.96	7.31+3.34†	6.17	6.84+16.9†	3.0	26.82+11.5	26.2	7.81+9.78	4.92
Trans	A+B	9.95+3.13	9.54	5.11+1.56	4.69	2.42+1.65	1.83	27.82+11.24	27.0	4.45+2.21	4.09
Trans w/o LN	A+B	10.8+3.54†	10.21	6.26+2.31†	5.83	3.55+2.48†	2.9	28.26+11.63	26.77	5.13+2.48†	4.46
Trans	A1+B	10.11+3.1	9.53	5.2+1.59	4.88	2.3+1.76	1.79	28.32+11.15	27.27	4.61+2.44	4.11
Trans w/o LN	A1+B	10.91+3.28†	10.1	6.21+1.79†	5.97	3.2+2.45†	2.49	28.74+10.75	27.15	5.49+3.23†	4.65
Trans	A2+B	10.4+3.08	10.04	5.52+1.55	5.19	2.94+1.37	2.61	28.02+10.85	27.0	5.13+2.39	4.66
Trans w/o LN	A2+B	11.13+3.74†	10.43	6.34+2.34†	5.62	3.44+2.54	2.65	28.33+11.62	27.02	6.4+3.85†	5.36

Table 16: Comparison between the Transformer with/without Layer Normalization. The HD95 mean and standard deviation on EMC test set(clinic C). † represents statistical significant difference($p<0.05$ between the annotated method and Transformer with Layer Normalization in the same training set.

		Mean		Prostate		Bladder		Rectum		SeminalVesicle	
Network	Training Set	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median
Trans	A1	85.49+3.64	86.64	86.89+3.11	87.27	94.81+1.52	95.0	86.41+4.86	87.54	73.85+11.06	77.25
Conv	A1	85.89+3.44†	87.02	87.13+3.41†	87.91	94.96+1.35	95.07	88.06+3.73†	88.81	73.42+12.69	77.69
Trans	A2	84.03+4.05	84.97	86.22+2.7	86.3	92.76+3.2	93.59	85.05+6.77	87.25	72.1+10.19	74.26
Conv	A2	84.02+4.16	85.48	86.42+2.71	86.3	93.78+2.05†	94.24	85.61+6.08†	87.63	70.28+11.7†	73.54
Trans	A1+B	84.98+3.42	86.04	86.33+2.63	86.59	94.24+5.84	95.16	85.46+4.43	86.53	73.89+10.05	76.93
Conv	A1+B	85.93+2.88†	86.8	86.73+2.99†	87.39	95.22+1.26†	95.33	86.58+4.27†	87.03	75.21+9.82†	77.91
Trans	A2+B	83.97+3.84	84.32	86.64+3.03	87.0	93.65+2.3	94.32	85.8+5.59	88.13	69.81+9.81	70.25
Conv	A2+B	84.87+3.32†	85.19	87.44+2.89†	87.95	94.38+1.87†	94.76	86.98+5.18†	89.24	70.69+8.89	72.36
Trans⋆	A1	85.9+3.29	86.81	87.02+3.14	87.4	95.15+1.32	95.33	87.0+4.76	87.87	74.42+10.59	78.33
Conv⋆	A1	86.19+3.09†	87.06	86.96+3.47	87.8	95.11+1.17	95.1	88.2+3.85†	89.0	74.48+10.78	77.72
Trans⋆	A2	85.22+3.53	85.56	86.94+2.77	87.36	94.19+1.73	94.35	85.98+5.94	87.82	73.77+8.71	74.82
Conv⋆	A2	84.98+3.69	86.19	86.79+2.82	87.05	94.5+1.69†	94.67	86.32+5.65†	88.07	72.3+9.61†	74.09

Table 17: Experiment 1: Replacing Swin-Transformer block with convolution block. The Dice mean and standard deviation validation set(clinic B). Training set A1 denotes the model is trained on A1 while validating on the A2, vice versa. ⋆ denotes the pretraining on clinic B and fintuning on Training data.† represents statistical significant difference($p<0.05$ between the annotated method and Trans with the same training set.

		Mean		Prostate		Bladder		Rectum		SeminalVesicle	
Network	Training Set	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median
Trans	A1	4.87+1.63	4.5	3.68+1.12	3.5	2.59+1.15	2.22	9.77+6.03	10.0	3.45+1.65	2.9
Conv	A1	4.38+1.4†	4.11	3.84+1.47	4.0	2.34+0.51†	2.17	7.95+4.81†	6.62	3.38+1.42	2.85
Trans	A2	6.7+3.18	6.5	3.84+0.99	3.96	5.78+7.99	2.58	13.09+10.73	7.95	4.09+2.16	3.4
Conv	A2	6.38+2.96	5.91	3.97+1.2	4.0	3.74+3.64†	2.28	13.06+10.19	8.61	4.76+2.76†	4.0
Trans	A1+B	5.14+1.48	4.9	3.96+1.06	4.0	2.85+1.91	2.32	10.04+5.37	8.06	3.7+1.8	3.21
Conv	A1+B	4.75+1.38†	4.52	3.84+1.11†	4.0	2.29+0.38†	2.17	9.48+5.27†	8.56	3.41+1.67†	2.79
Trans	A2+B	5.26+2.45	4.41	3.72+0.95	3.66	2.81+1.58	2.38	9.89+8.83	6.05	4.64+2.29	4.0
Conv	A2+B	5.36+2.45	4.75	3.69+1.1	3.48	2.33+0.51†	2.15	10.65+9.06	7.25	4.77+2.4	4.08
Trans⋆	A1	4.75+1.79	4.13	3.62+1.02	3.9	2.46+1.14	2.15	9.59+6.73	8.0	3.34+1.48	2.79
Conv⋆	A1	4.44+1.63†	4.01	3.87+1.41†	4.0	2.22+0.28	2.15	8.23+5.43†	6.06	3.43+1.87	2.8
Trans⋆	A2	5.45+2.61	4.47	3.62+0.99	3.47	2.61+0.86	2.35	11.84+9.35	8.2	3.74+1.63	3.25
Conv⋆	A2	5.77+3.12	4.72	3.88+1.13†	3.99	3.2+6.92†	2.0	11.8+9.28	8.05	4.19+1.99†	4.0

Table 18: Experiment 1: Replacing Swin-Transformer block with convolution block. The HD95 mean and standard deviation validation set(clinic B). Training set A1 denotes the model is trained on A1 while validating on the A2, vice versa. ⋆ denotes the pretraining on clinic B and fintuning on Training data.† represents statistical significant difference($p<0.05$ between the annotated method and Trans with the same training set.

		Mean		Prostate		Bladder		Rectum		SeminalVesicle	
Network	Training Set	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median
Trans	A1	85.49+3.56	86.52	86.87+3.09	87.41	94.94+1.48	95.17	86.52+4.67	87.1	73.61+11.12	77.86
Conv	A1	85.87+3.38†	86.93	87.11+3.21†	87.6	95.03+1.32	95.13	87.85+3.86†	88.73	73.51+12.06	77.16
Trans	A2	84.18+4.01	85.06	86.39+2.74	86.79	92.85+3.13	93.91	85.17+6.49	87.15	72.33+10.63	75.1
Conv	A2	84.33+3.89	85.54	86.41+2.8	86.54	93.77+1.86†	94.19	85.64+5.95†	87.33	71.49+10.36†	74.85
Trans	A1+B	83.45+3.9	84.53	84.54+2.88	84.72	93.59+6.7	94.9	83.64+4.9	84.25	72.05+10.45	75.12
Conv	A1+B	85.69+3.05†	86.69	86.75+2.82†	87.0	95.0+1.8†	95.25	86.18+4.43†	86.97	74.85+10.2†	78.34
Trans	A2+B	83.18+4.21	83.61	85.93+3.23	86.37	93.1+2.79	93.92	85.31+5.58	87.51	68.39+10.8	70.16
Conv	A2+B	84.64+3.31†	84.84	87.27+2.93†	87.63	94.23+1.89†	94.53	86.8+5.1†	89.05	70.26+9.31†	71.79
Trans⋆	86.27+3.02†	87.28	87.15+3.1	87.67	95.17+1.16	95.18	88.1+4.02†	88.6	74.66+10.74†	77.6	
Conv⋆	A1	85.7+3.21	86.79	86.91+2.98	87.46	95.12+1.42	95.33	86.69+4.49	87.3	74.06+10.44	77.65
Trans⋆	A2	84.79+3.92	85.38	86.66+2.74	86.96	93.95+1.8	94.21	85.55+6.09	87.74	73.0+10.19	74.69
Conv⋆	A2	84.97+3.76	86.06	86.76+2.67	87.07	94.42+1.69†	94.7	86.16+5.88†	88.26	72.54+9.72	75.35

Table 19: Experiment 1: Replacing Swin-Transformer block with convolution block after halving the parameters. The Dice mean and standard deviation validation set(clinic B). Training set A1 denotes the model is trained on A1 while validating on the A2, vice versa. ⋆ denotes the pretraining on clinic B and fintuning on Training data.† represents statistical significant difference(p<0.05 between the annotated method and Trans with the same training set.

		Mean		Prostate		Bladder		Rectum		SeminalVesicle	
Network	Training Set	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median
Trans	A1	4.84+1.49	4.53	3.68+1.12	3.6	2.51+0.94	2.21	9.72+5.75	8.0	3.46+1.59	2.87
Conv	A1	4.46+1.5†	4.15	3.85+1.4	4.0	2.27+0.4†	2.16	8.31+5.04†	7.22	3.42+1.45	2.91
Trans	A2	6.33+2.8	6.06	3.75+0.86	3.9	5.36+7.29	2.61	12.15+9.39	8.02	4.08+2.29	3.33
Conv	A2	6.23+2.9	5.35	3.94+1.22	4.0	3.6+3.3†	2.38	12.99+10.29	8.34	4.38+2.4†	3.77
Trans	A1+B	5.69+1.63	5.34	4.6+1.17	4.0	3.41+2.51	2.69	10.87+5.5	10.0	3.88+1.72	3.46
Conv	A1+B	4.89+1.41†	4.73	3.87+1.09†	4.0	2.53+1.35†	2.19	9.54+5.26†	8.82	3.62+2.0†	2.98
Trans	A2+B	5.37+2.43	4.47	3.81+0.93	3.91	3.06+1.87	2.6	9.82+8.6	6.27	4.79+2.47	4.0
Conv	A2+B	5.53+2.5	4.82	3.71+1.02	3.61	2.38+0.56†	2.18	11.04+9.54†	7.55	4.99+2.49†	4.18
Trans⋆	A1	4.83+1.65	4.51	3.65+0.97	3.82	2.45+0.93	2.16	9.81+6.21	8.0	3.43+1.67	2.81
Conv⋆	A1	4.36+1.53†	3.91	3.81+1.36	3.85	2.22+0.31†	2.16	8.05+5.29†	6.11	3.38+1.86	2.79
Trans⋆	A2	5.75+2.75	4.8	3.66+0.91	3.5	2.89+1.34	2.54	12.4+9.97	8.26	4.05+2.35	3.13
Conv⋆	A2	5.59+2.82	4.4	3.8+1.11	3.67	2.66+1.42†	2.17	11.65+9.48	7.94	4.23+2.41†	3.86

Table 20: Experiment 1: Replacing Swin-Transformer block with convolution block after halving the parameters. The HD95 mean and standard deviation validation set(clinic B). Training set A1 denotes the model is trained on A1 while validating on the A2, vice versa. ⋆ denotes the pretraining on clinic B and fintuning on Training data.† represents statistical significant difference(p<0.05 between the annotated method and Trans with the same training set.

		Mean		Prostate		Bladder		Rectum		SeminalVesicle	
Network	Training Set	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median
Trans	A1	85.49+3.64	86.64	86.89+3.11	87.27	94.81+1.52	95.0	86.41+4.86	87.54	73.85+11.06	77.25
Pool	A1	85.62+3.62	86.82	87.21+3.05†	87.74	94.72+1.59	94.98	86.38+4.93	87.41	74.18+10.75	78.0
Trans	A2	84.03+4.05	84.97	86.22+2.7	86.3	92.76+3.2	93.59	85.05+6.77	87.25	72.1+10.19	74.26
Pool	A2	83.87+3.91	84.55	85.6+3.1†	86.17	92.78+3.06	93.82	85.28+6.01	87.68	71.8+9.56	73.66
Trans	A1+B	84.98+3.42	86.04	86.33+2.63	86.59	94.24+5.84	95.16	85.46+4.43	86.53	73.89+10.05	76.93
Pool	A1+B	84.91+3.67†	86.25	86.5+2.81†	86.92	93.61+6.27†	94.96	85.91+4.75†	87.15	73.6+9.97	76.15
Trans	A2+B	82.94+4.45	83.71	85.57+3.5	86.15	92.4+2.96	93.21	83.93+6.38	86.42	69.84+10.94	72.22
Pool	A2+B	83.63+4.0†	84.17	86.36+3.26†	86.94	93.42+2.36†	93.76	85.74+5.74†	88.47	68.99+10.17	71.9
Trans⋆	A1	85.9+3.29	86.81	87.02+3.14	87.4	95.15+1.32	95.33	87.0+4.76	87.87	74.42+10.59	78.33
Pool⋆	A1	85.77+3.37	86.77	87.02+2.86	87.45	94.89+1.46†	95.18	87.07+4.28	87.83	74.09+10.36†	77.85
Trans⋆	A2	85.22+3.53	85.56	86.94+2.77	87.36	94.19+1.73	94.35	85.98+5.94	87.82	73.77+8.71	74.82
Pool⋆	A2	84.21+3.47†	84.53	85.8+3.02†	86.17	92.64+3.55†	93.71	85.24+6.13†	87.29	73.15+7.83	74.9

Table 21: Experiment 2: Replacing the Self-Attention with Pooling: The Dice mean and standard deviation validation set(clinic B). Training set A1 denotes the model is trained on A1 while validating on the A2, vice versa. ⋆ denotes the pretraining on clinic B and fintuning on Training data.† represents statistical significant difference(p<0.05 between the annotated method and Trans with the same training set.

		Mean		Prostate		Bladder		Rectum		SeminalVesicle	
Network	Training Set	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median
Trans	A1	4.87+1.63	4.5	3.68+1.12	3.5	2.59+1.15	2.22	9.77+6.03	10.0	3.45+1.65	2.9
Pool	A1	4.81+1.66	4.41	3.57+0.98†	3.84	2.55+0.97	2.22	9.55+5.87	8.12	3.56+1.76	2.81
Trans	A2	6.7+3.18	6.5	3.84+0.99	3.96	5.78+7.99	2.58	13.09+10.73	7.95	4.09+2.16	3.4
Pool	A2	6.7+3.1	6.32	4.04+1.1†	4.0	5.16+6.27	2.67	13.38+10.56	8.47	4.22+2.26	3.38
Trans	A1+B	5.14+1.48	4.9	3.96+1.06	4.0	2.85+1.91	2.32	10.04+5.37	8.06	3.7+1.8	3.21
Pool	A1+B	4.91+1.57†	4.48	3.86+1.04†	4.0	2.9+2.01	2.25	9.2+5.2†	8.0	3.67+1.83	3.08
Trans	A2+B	5.58+2.5	4.57	3.9+0.92	4.0	3.52+2.2	2.74	10.5+9.03	6.0	4.42+2.47	3.52
Pool	A2+B	5.86+2.94	4.86	3.97+1.4	3.64	2.83+0.91†	2.59	11.52+10.49	6.95	5.1+2.43†	4.08
Trans⋆	A1	4.75+1.79	4.13	3.62+1.02	3.9	2.46+1.14	2.15	9.59+6.73	8.0	3.34+1.48	2.79
Pool⋆	A1	4.49+1.45†	4.09	3.66+0.98	3.68	2.47+0.7†	2.19	8.32+5.14†	6.68	3.51+1.62†	2.96
Trans⋆	A2	5.45+2.61	4.47	3.62+0.99	3.47	2.61+0.86	2.35	11.84+9.35	8.2	3.74+1.63	3.25
Pool⋆	A2	6.73+4.21†	6.24	3.88+0.91†	4.0	6.67+14.53†	2.67	12.52+9.57†	8.36	3.85+1.65	3.33

Table 22: Experiment 2: Replacing the Self-Attention with Pooling: The HD95 mean and standard deviation validation set(clinic B). Training set A1 denotes the model is trained on A1 while validating on the A2, vice versa. ⋆ denotes the pre-training on clinic B and fintuning on Training data.† represents statistical significant difference(p<0.05 between the annotated method and Trans with the same training set.

		Mean		Prostate		Bladder		Rectum		SeminalVesicle	
Network	Training Set	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median
AvgPooling	A1	85.41+3.69	86.75	87.01+3.14	87.56	94.74+1.65	94.93	86.68+5.15†	87.53	73.2+11.09†	76.62
MaxPooling	A1	85.62+3.62	86.82	87.21+3.05	87.74	94.72+1.59	94.98	86.38+4.93	87.41	74.18+10.75	78.0
AvgPooling	A2	84.26+3.77†	85.25	86.27+2.76†	86.72	92.41+3.44†	93.68	85.5+6.11†	87.21	72.89+9.2†	75.01
MaxPooling	A2	83.87+3.91	84.55	85.6+3.1	86.17	92.78+3.06	93.82	85.28+6.01	87.68	71.8+9.56	73.66

Table 23: Comparison between Avg-pooling and Max-pooling: The Dice mean and standard deviation on validation set. Training set A1 denotes the model is trained on A1 while validating on the A2, vice versa. † represents statistical significant difference(p<0.05 between the annotated method and Trans(Max-pooling) with the same training set

		Mean		Prostate		Bladder		Rectum		SeminalVesicle	
Network	Training Set	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median
AvgPooling	A1	4.72+1.59	4.4	3.73+1.24†	3.77	2.49+0.87	2.19	8.96+5.69†	8.0	3.68+1.63†	3.13
MaxPooling	A1	4.81+1.66	4.41	3.57+0.98	3.84	2.55+0.97	2.22	9.55+5.87	8.12	3.56+1.76	2.81
AvgPooling	A2	6.2+2.69†	5.98	3.75+0.89†	3.86	5.67+7.42†	2.91	11.5+8.63†	8.31	3.87+1.84†	3.35
MaxPooling	A2	6.7+3.1	6.32	4.04+1.1	4.0	5.16+6.27	2.67	13.38+10.56	8.47	4.22+2.26	3.38

Table 24: Comparison between Avg-pooling and Max-pooling: The HD95 mean and standard deviation on validation set(clinic B). Training set A1 denotes the model is trained on A1 while validating on the A2, vice versa. † represents statistical significant difference(p<0.05 between the annotated method and Trans(Max-pooling) with the same training set

		Mean		Prostate		Bladder		Rectum		SeminalVesicle	
Network	Training Set	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median
Trans w/o	A1	85.58+3.69	86.75	87.06+3.17	87.42	94.81+1.44	94.89	86.7+4.78	87.7	73.77+11.32	77.32
TransR	A1	85.49+3.64†	86.64	86.89+3.11†	87.27	94.81+1.52	95.0	86.41+4.86†	87.54	73.85+11.06	77.25
TransA	A1	85.74+3.5†‡	86.79	87.07+2.98	87.3	94.87+1.41†	95.04	87.08+4.38†‡	87.86	73.94+11.32	77.9
Trans w/o	A2	84.13+3.99	84.99	86.02+2.97	86.53	92.56+3.71	93.73	85.21+6.69	87.49	72.73+9.69	74.83
TransR	A2	84.03+4.05†	84.97	86.22+2.7	86.3	92.76+3.2	93.59	85.05+6.77†	87.25	72.1+10.19†	74.26
TransA	A2	84.3+3.99†‡	85.39	86.16+2.96	86.53	92.77+3.27†‡	93.75	85.26+6.75‡	87.63	72.99+9.57‡	74.98
Trans w/o	A1+B	84.59+3.42	85.54	85.79+2.87	86.2	94.29+4.3	95.04	84.82+4.76	85.93	73.45+10.21	76.44
TransR	A1+B	84.98+3.42†	86.04	86.33+2.63†	86.59	94.24+5.84†	95.16	85.46+4.43†	86.53	73.89+10.05†	76.93
TransA	A1+B	83.94+4.0†‡	85.24	85.19+2.96†‡	85.48	93.55+7.62†‡	94.81	83.77+5.32†‡	84.86	73.24+10.11‡	75.44
Trans w/o	A2+B	84.3+3.57	84.69	86.7+3.12	86.9	93.75+2.3	94.31	86.04+5.48	88.33	70.71+9.12	72.51
TransR	A2+B	83.97+3.84†	84.32	86.64+3.03	87.0	93.65+2.3†	94.32	85.8+5.59†	88.13	69.81+9.81	70.25
TransA	A2+B	84.18+3.54†‡	84.35	86.51+3.17	86.6	93.54+2.28†‡	94.09	85.93+5.58†	88.26	70.75+8.71‡	72.33
Trans w/o ⋆	A1	85.86+3.3	86.95	87.08+3.02	87.41	95.03+1.32	95.15	86.89+4.5	87.45	74.42+10.58	78.07
TransR⋆	A1	85.9+3.29	86.81	87.02+3.14	87.4	95.15+1.32†	95.33	87.0+4.76	87.87	74.42+10.59	78.33
TransA⋆	A1	85.94+3.33†	87.06	87.1+3.03	87.58	95.12+1.26†‡	95.3	86.94+4.39	87.58	74.59+10.64†	78.43
Trans w/o ⋆	A2	85.18+3.54	85.63	86.99+2.65	87.23	94.14+1.73	94.32	85.83+5.82	87.52	73.76+9.13	74.96
TransR⋆	A2	85.22+3.53	85.56	86.94+2.77	87.36	94.19+1.73†	94.35	85.98+5.94†	87.82	73.77+8.71	74.82
TransA⋆	A2	84.97+3.81†‡	85.54	86.77+2.67†‡	87.16	94.22+1.72†	94.34	85.78+6.08‡	88.03	73.11+10.05†	74.64

Table 25: Experiment 3-Evaluating Positional Encoding: The Dice mean and standard deviation on validation set(clinic B). Training set A1 denotes the model is trained on A1 while validating on the A2, vice versa. † represents statistical significant difference($p<0.05$ between the annotated method and Transformer without Positional encoding with the same training set. ‡ represents statistical significant difference($p<0.05$ between the annotated method and Transformer with relative positional bias(Trans) with the same training set.

		Mean		Prostate		Bladder		Rectum		SeminalVesicle	
Network	Training Set	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median
Trans w/o	A1	4.79+1.62	4.53	3.65+1.16	3.39	2.58+1.04	2.22	9.51+5.85	8.0	3.43+1.56	2.77
TransR	A1	4.87+1.63†	4.5	3.68+1.12	3.5	2.59+1.15	2.22	9.77+6.03	10.0	3.45+1.65	2.9
TransA	A1	4.7+1.56†‡	4.45	3.67+1.11	3.65	2.53+1.02	2.22	9.18+5.77†‡	8.0	3.42+1.7	2.79
Trans w/o	A2	6.56+3.06	6.45	3.89+1.07	4.0	5.85+8.2	2.65	12.5+9.74	8.15	3.99+2.06	3.3
TransR	A2	6.7+3.18	6.5	3.84+0.99	3.96	5.78+7.99	2.58	13.09+10.73†	7.95	4.09+2.16	3.4
TransA	A2	6.69+3.42‡	6.39	3.81+1.05†	3.73	6.55+10.65	2.5	12.49+9.95	8.0	3.92+1.99‡	3.26
Trans w/o	A1+B	5.48+1.91	5.15	4.11+1.08	4.0	2.84+1.34	2.34	11.05+5.85	10.0	3.93+3.48	3.0
TransR	A1+B	5.14+1.48†	4.9	3.96+1.06†	4.0	2.85+1.91	2.32	10.04+5.37†	8.06	3.7+1.8	3.21
TransA	A1+B	5.74+1.91†‡	5.16	4.34+1.14†‡	4.0	3.32+3.18†‡	2.49	11.53+6.72‡	10.0	3.76+1.86	3.11
Trans w/o	A2+B	5.07+2.26	4.23	3.65+0.86	3.56	2.58+0.73	2.3	9.63+8.31	6.0	4.43+2.08	3.95
TransR	A2+B	5.26+2.45	4.41	3.72+0.95†	3.66	2.81+1.58†	2.38	9.89+8.83	6.05	4.64+2.29‡	4.0
TransA	A2+B	5.06+2.15	4.48	3.74+0.89†	3.95	2.78+0.89†	2.44	9.32+7.7	6.03	4.4+1.96	3.96
Trans w/o ⋆	A1	4.84+1.83	4.33	3.62+1.02	3.77	2.38+0.74	2.17	9.96+6.83	8.0	3.39+1.75	2.78
TransR⋆	A1	4.75+1.79	4.13	3.62+1.02†	3.9	2.46+1.14	2.15	9.59+6.73	8.0	3.34+1.48†	2.79
TransA⋆	A1	4.74+1.61	4.32	3.55+0.95†	3.71	2.36+0.72	2.17	9.72+6.03	8.0	3.34+1.65†	2.76
Trans w/o ⋆	A2	5.52+2.59	4.56	3.58+0.92	3.42	2.69+0.99	2.36	12.06+9.45	8.15	3.75+1.73	3.14
TransR⋆	A2	5.45+2.61	4.47	3.62+0.99	3.47	2.61+0.86	2.35	11.84+9.35	8.2	3.74+1.63	3.25
TransA⋆	A2	5.65+2.7‡	4.49	3.67+0.91†	3.55	2.57+0.74†	2.32	12.39+9.58‡	8.02	3.98+2.14†‡	3.22

Table 26: Experiment 3:Evaluating Positional Encoding. The HD95 mean and standard deviation on validation set(clinic B). Training set A1 denotes the model is trained on A1 while validating on the A2, vice versa. † represents statistical significant difference($p<0.05$ between the annotated method and Transformer without Positional encoding with the same training set. ‡ represents statistical significant difference($p<0.05$ between the annotated method and Transformer with relative positional bias(Trans) with the same training set.

		Mean		Prostate		Bladder		Rectum		SeminalVesicle	
Network	Training Set	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median
TransA	A1	85.74+3.5†	86.79	87.07+2.98	87.3	94.87+1.41	95.04	87.08+4.38	87.86	73.94+11.32†	77.9
TransS	A1	85.52+3.77	86.72	87.13+3.0	87.49	94.84+1.52	95.11	86.81+4.83	87.72	73.28+12.18	77.84
TransA	A2	84.3+3.99†	85.39	86.16+2.96	86.53	92.77+3.27	93.75	85.26+6.75	87.63	72.99+9.57†	74.98
TransS	A2	84.14+4.2	85.07	86.08+3.1	86.37	92.67+3.51	93.65	85.14+6.85	87.44	72.65+10.33	75.52
TransA	A1+B	83.94+4.0	85.24	85.19+2.96†	85.48	93.55+7.62†	94.81	83.77+5.32	84.86	73.24+10.11	75.44
TransS	A1+B	84.08+3.71	85.18	85.53+2.54	85.92	94.04+4.87	94.98	83.79+5.55	84.7	72.95+10.56	75.96
TransA	A2+B	84.18+3.54†	84.35	86.51+3.17†	86.6	93.54+2.28†	94.09	85.93+5.58†	88.26	70.75+8.71†	72.33
TransS	A2+B	83.79+3.75	83.98	86.17+3.33	86.4	93.44+2.32	94.04	85.61+5.57	88.0	69.94+9.75	71.07
TransA⋆	A1	85.94+3.33	87.06	87.1+3.03†	87.58	95.12+1.26	95.3	86.94+4.39†	87.58	74.59+10.64	78.43
TransS⋆	A1	85.93+3.22	87.02	86.92+3.02	87.14	95.09+1.29	95.26	87.18+4.37	87.79	74.52+10.61	78.41
TransA⋆	A2	84.97+3.81	85.54	86.77+2.67	87.16	94.22+1.72†	94.34	85.78+6.08	88.03	73.11+10.05	74.64
TransS⋆	A2	85.0+3.67	85.4	86.72+2.78	87.02	94.08+1.78	94.35	85.75+6.05	87.48	73.45+9.12	74.18

Table 27: Comparison between learned and unlearned sinusoid absolute position embedding. The Dice mean and standard deviation on validation set(clinic B). Training set A1 denotes the model is trained on A1 while validating on the A2, vice versa. ⋆ denotes the pretraining on clinic B and finetuning on clinic A .† represents statistical significant difference($p<0.05$ between the annotated method and Trans with absolute position embedding in the same training set.

		Mean		Prostate		Bladder		Rectum		SeminalVesicle	
Network	Training Set	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median
TransA	A1	4.7+1.56†	4.45	3.67+1.11	3.65	2.53+1.02	2.22	9.18+5.77	8.0	3.42+1.7†	2.79
TransS	A1	4.79+1.65	4.39	3.68+1.2	4.0	2.58+1.13	2.22	9.38+6.02	8.0	3.53+1.87	2.95
TransA	A2	6.69+3.42	6.39	3.81+1.05†	3.73	6.55+10.65	2.5	12.49+9.95	8.0	3.92+1.99†	3.26
TransS	A2	6.56+3.04	6.47	3.93+1.25	4.0	5.84+8.45	2.63	12.46+9.87	8.1	4.03+2.1	3.33
TransA	A1+B	5.74+1.91	5.16	4.34+1.14†	4.0	3.32+3.18†	2.49	11.53+6.72	10.0	3.76+1.86	3.11
TransS	A1+B	5.76+1.76	5.42	4.22+1.12	4.0	3.0+1.61	2.41	11.95+6.64	12.0	3.88+1.89	3.33
TransA	A2+B	5.06+2.15†	4.48	3.74+0.89†	3.95	2.78+0.89	2.44	9.32+7.7†	6.03	4.4+1.96	3.96
TransS	A2+B	5.18+2.24	4.43	3.83+0.95	3.97	2.7+0.78	2.38	9.63+8.21	6.0	4.56+2.34	4.0
TransA⋆	A1	4.74+1.61	4.32	3.55+0.95†	3.71	2.36+0.72†	2.17	9.72+6.03	8.0	3.34+1.65	2.76
TransS⋆	A1	4.75+1.77	4.27	3.67+1.02	3.88	2.41+0.78	2.18	9.59+6.63	8.0	3.32+1.57	2.75
TransA⋆	A2	5.65+2.7	4.49	3.67+0.91	3.55	2.57+0.74†	2.32	12.39+9.58	8.02	3.98+2.14	3.22
TransS⋆	A2	5.69+2.64	4.53	3.73+0.96	3.74	3.04+3.25	2.51	12.11+9.36	8.11	3.88+1.93	3.26

Table 28: Comparison between learned and unlearned sinusoid absolute position embedding. The HD95 mean and standard deviation on validation set(clinic B). Training set A1 denotes the model is trained on A1 while validating on the A2, vice versa. ⋆ denotes the pretraining on clinic B and finetuning on clinic A .† represents statistical significant difference($p<0.05$ between the annotated method and Trans with absolute position embedding in the same training set.

		Mean		Prostate		Bladder		Rectum		SeminalVesicle	
Network	Training Set	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median
Trans	A2	84.03+4.05	84.97	86.22+2.7	86.3	92.76+3.2	93.59	85.05+6.77	87.25	72.1+10.19	74.26
Conv	A2	84.02+4.16	85.48	86.42+2.71	86.3	93.78+2.05	94.24	85.61+6.08	87.63	70.28+11.7	73.54
Trans-L	A2	84.55+3.77†‡	85.24	86.41+2.88†	86.65	92.86+3.41†‡	93.82	85.33+6.76†	87.68	73.58+8.92†‡	75.3
Trans-S	A2	84.18+4.01	85.06	86.39+2.74†	86.79	92.85+3.13†‡	93.91	85.17+6.49‡	87.15	72.33+10.63‡	75.1
Trans	A2+B	83.97+3.84	84.32	86.64+3.03	87.0	93.65+2.3	94.32	85.8+5.59	88.13	69.81+9.81	70.25
Conv	A2+B	84.87+3.32	85.19	87.44+2.89	87.95	94.38+1.87	94.76	86.98+5.18	89.24	70.69+8.89	72.36
Trans-L	A2+B	84.48+3.6†‡	84.62	87.01+3.04†‡	87.43	94.05+2.1†‡	94.46	86.45+5.49†‡	88.26	70.41+9.29†	71.21
Trans-S	A2+B	83.18+4.21†‡	83.61	85.93+3.23†‡	86.37	93.1+2.79†‡	93.92	85.31+5.58†‡	87.51	68.39+10.8†‡	70.16
Trans	A1	85.49+3.64	86.64	86.89+3.11	87.27	94.81+1.52	95.0	86.41+4.86	87.54	73.85+11.06	77.25
Conv	A1	85.89+3.44	87.02	87.13+3.41	87.91	94.96+1.35	95.07	88.06+3.73	88.81	73.42+12.69	77.69
Trans-L	A1	86.06+3.36†	87.09	87.25+3.09†	87.62	95.07+1.4†‡	95.23	87.44+4.3†‡	88.38	74.47+10.71†	78.07
Trans-S	A1	85.49+3.56‡	86.52	86.87+3.09‡	87.41	94.94+1.48†	95.17	86.52+4.67‡	87.1	73.61+11.12	77.86
Trans	A1+B	84.98+3.42	86.04	86.33+2.63	86.59	94.24+5.84	95.16	85.46+4.43	86.53	73.89+10.05	76.93
Conv	A1+B	85.93+2.88	86.8	86.73+2.99	87.39	95.22+1.26	95.33	86.58+4.27	87.03	75.21+9.82	77.91
Trans-L	A1+B	85.31+3.3†‡	86.32	86.6+2.8†	86.93	94.56+3.42‡	95.11	85.74+4.8†‡	86.52	74.33+10.24†‡	77.45
Trans-S	A1+B	83.45+3.9‡†	84.53	84.54+2.88‡†	84.72	93.59+6.7‡†	94.9	83.64+4.9	84.25‡†	72.05+10.45‡†	75.12

Table 29: Comparison between models of different size. The Dice mean and standard deviation on validation set(clinic B). Training set A1 denotes the model is trained on A1 while validating on the A2, vice versa. † represents statistical significant difference($p<0.05$ between the annotated method and Transformer with the same training set. ‡ represents statistical significant difference($p<0.05$ between the annotated method and the convolution with the same training set.

		Mean		Prostate		Bladder		Rectum		SeminalVesicle	
Network	Training Set	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median
Trans	A2	6.7+3.18	6.5	3.84+0.99	3.96	5.78+7.99	2.58	13.09+10.73	7.95	4.09+2.16	3.4
Conv	A2	6.38+2.96	5.91	3.97+1.2	4.0	3.74+3.64	2.28	13.06+10.19	8.61	4.76+2.76	4.0
Trans-L	A2	6.56+3.2†	6.4	3.73+0.9†‡	3.8	5.75+8.1‡	2.47	12.99+10.4	8.24	3.76+1.75†‡	3.25
Trans-S	A2	6.33+2.8†	6.06	3.75+0.86†‡	3.9	5.36+7.29‡	2.61	12.15+9.39†‡	8.02	4.08+2.29‡	3.33
Trans	A2+B	5.26+2.45	4.41	3.72+0.95	3.66	2.81+1.58	2.38	9.89+8.83	6.05	4.64+2.29	4.0
Conv	A2+B	5.36+2.45	4.75	3.69+1.1	3.48	2.33+0.51	2.15	10.65+9.06	7.25	4.77+2.4	4.08
Trans-L	A2+B	5.21+2.37	4.5	3.64+0.86	3.68	2.45+0.55†‡	2.19	10.14+8.68	6.19	4.62+2.22	4.0
Trans-S	A2+B	5.37+2.43†	4.47	3.81+0.93†‡	3.91	3.06+1.87†‡	2.6	9.82+8.6	6.27	4.79+2.47†	4.0
Trans	A1	4.87+1.63	4.5	3.68+1.12	3.5	2.59+1.15	2.22	9.77+6.03	10.0	3.45+1.65	2.9
Conv	A1	4.38+1.4	4.11	3.84+1.47	4.0	2.34+0.51	2.17	7.95+4.81	6.62	3.38+1.42	2.85
Trans-L	A1	4.63+1.6†	4.24	3.62+1.2†	3.63	2.41+0.78†	2.18	9.17+5.96†	8.0	3.34+1.51†	2.78
Trans-S	A1	4.84+1.49‡	4.53	3.68+1.12	3.6	2.51+0.94†	2.21	9.72+5.75‡	8.0	3.46+1.59	2.87
Trans	A1+B	5.14+1.48	4.9	3.96+1.06	4.0	2.85+1.91	2.32	10.04+5.37	8.06	3.7+1.8	3.21
Conv	A1+B	4.75+1.38	4.52	3.84+1.11	4.0	2.29+0.38	2.17	9.48+5.27	8.56	3.41+1.67	2.79
Trans-L	A1+B	5.1+1.65‡	4.81	3.86+1.12†	4.0	2.71+1.68†‡	2.22	10.24+5.71‡	8.76	3.59+2.17†‡	2.87
Trans-S	A1+B	5.69+1.63‡†	5.34	4.6+1.17‡†	4.0	3.41+2.51‡†	2.69	10.87+5.5‡†	10.0	3.88+1.72‡†	3.46

Table 30: Comparison between models of different size. The HD95 mean and standard deviation on validation set(clinic B). Training set A1 denotes the model is trained on A1 while validating on the A2, vice versa. † represents statistical significant difference(p<0.05 between the annotated method and Transformer with the same training set. ‡ represents statistical significant difference(p<0.05 between the annotated method and the convolution with the same training set.

		Mean		Prostate		Bladder		Rectum		SeminalVesicle	
Network	Training Set	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median
Trans	A1	85.49+3.64	86.64	86.89+3.11	87.27	94.81+1.52	95.0	86.41+4.86	87.54	73.85+11.06	77.25
Trans w/o LN	A1	84.74+3.87†	85.83	86.16+3.1†	86.5	94.35+1.65†	94.76	85.48+5.61†	86.53	72.96+10.94†	76.69
Trans	A2	84.03+4.05	84.97	86.22+2.7	86.3	92.76+3.2	93.59	85.05+6.77	87.25	72.1+10.19	74.26
Trans w/o LN	A2	82.73+4.61†	84.03	84.59+3.39†	85.48	92.23+3.25†	93.59	84.01+7.25†	86.62	70.08+11.07†	71.11
Trans	A1+B	84.98+3.42	86.04	86.33+2.63	86.59	94.24+5.84	95.16	85.46+4.43	86.53	73.89+10.05	76.93
Trans w/o LN	A1+B	82.13+4.88†	83.75	84.35+3.43†	84.86	92.43+10.4†	94.81	81.83+6.03†	82.91	69.91+11.59†	72.52
Trans	A2+B	82.94+4.45	83.71	85.57+3.5	86.15	92.4+2.96	93.21	83.93+6.38	86.42	69.84+10.94	72.22
Trans w/o LN	A2+B	81.3+5.4†	82.02	84.28+3.47†	84.48	91.02+10.1†	93.47	84.17+5.67	86.03	65.74+11.55†	66.05

Table 31: Comparison between the Transformer with/without Layer Normalization. The HD95 mean and standard deviation on validation set(clinic B). Training set A1 denotes the model is trained on A1 while validating on the A2, vice versa. † represents statistical significant difference(p<0.05 between the annotated method and Transformer with Layer Normalizationin the same training set.

		Mean		Prostate		Bladder		Rectum		SeminalVesicle	
Network	Training Set	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median	$\mu + \sigma$	median
Trans	A1	4.87+1.63	4.5	3.68+1.12	3.5	2.59+1.15	2.22	9.77+6.03	10.0	3.45+1.65	2.9
Trans w/o LN	A1	5.06+1.74	4.72	3.92+1.08†	4.0	2.88+1.57†	2.41	9.56+5.84	8.0	3.88+2.41†	3.11
Trans	A2	6.7+3.18	6.5	3.84+0.99	3.96	5.78+7.99	2.58	13.09+10.73	7.95	4.09+2.16	3.4
Trans w/o LN	A2	6.92+3.17†	6.31	4.32+1.15†	4.0	5.59+6.85†	2.73	13.27+10.86	8.65	4.5+2.34†	3.47
Trans	A1+B	5.14+1.48	4.9	3.96+1.06	4.0	2.85+1.91	2.32	10.04+5.37	8.06	3.7+1.8	3.21
Trans w/o LN	A1+B	7.05+4.27	6.15	4.76+1.28	4.39	3.75+5.74	2.62	15.41+16.0	12.0	4.28+2.16	4.0
Trans	A2+B	5.26+2.45†	4.41	3.72+0.95†	3.66	2.81+1.58†	2.38	9.89+8.83†	6.05	4.64+2.29†	4.0
Trans w/o LN	A2+B	5.91+2.4†	5.2	4.36+1.16†	4.0	3.54+1.99†	3.08	10.3+7.98	6.28	5.43+3.11†	4.2

Table 32: Comparison between the Transformer with/without Layer Normalization. The HD95 mean and standard deviation on validation set(clinic B). Training set A1 denotes the model is trained on A1 while validating on the A2, vice versa. † represents statistical significant difference(p<0.05 between the annotated method and Transformer with Layer Normalizationin the same training set.

3 Supplements

In this project, we analyze the performance of the window-based Transformer based on the nnFormer[22] that is a 3D version Swin-Transformer. Therefore, we will first briefly introduce the Vision Transformer, Swin-Transformer block[3]. We later discuss about the modification nnFormer[22] to adapt it to 3D medical segmentation. Finally, we introduce the prostate data set we use in our project.

3.1. Vision Transformer

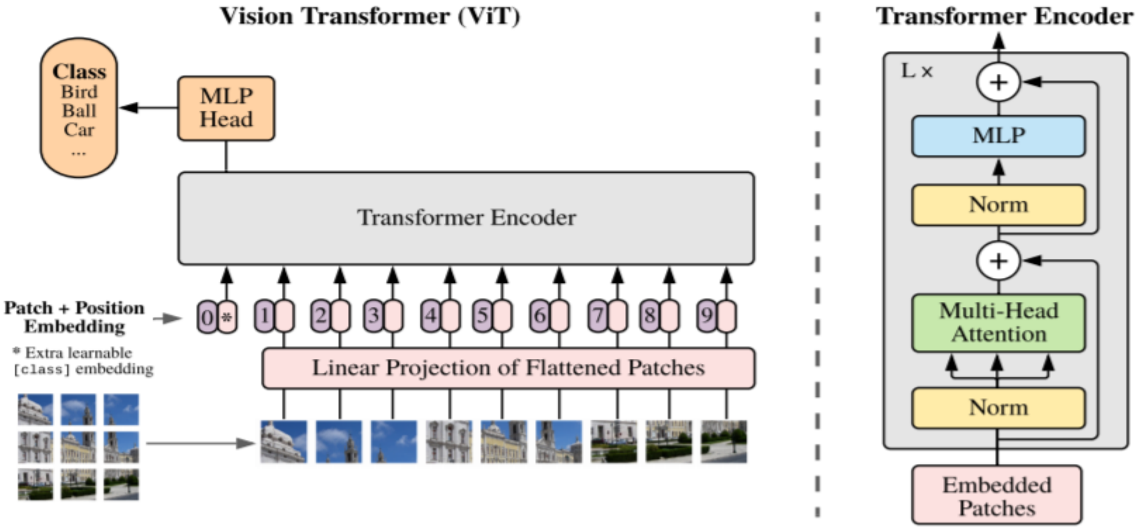


Figure 3.1: The overview of Vision Transformer structure and a Transformer encoder block. The picture is adapted from the ViT paper[8].

As shown in figure 3.1, the vanilla ViT, consisting of multiple Transformer block, is first designed for image classification tasks. The input images are first divided into non-overlapped patches and then flattened to suffice the Transformer input requirements. A input image $x \in \mathbb{R}^{H \times W \times C}$ with height H, width W, channel C is first split to N patches $p \in \mathbb{R}^{P \times P \times C}$, where P is the patch size, $N = \frac{HW}{P2}$ is the number of patches. After the division, the image is flatten to a series of patches $x \in \mathbb{R}^{N \times (P^2 \cdot C)}$. The flattened patches are then forwarded to the Linear Projection layer which maps the channel dimension C to arbitrary dimension D. Finally, a series of patch embeddings with channel dimension D are added with position embeddings and fed to the stacked Transformer blocks. To complete the image classification task, a class embedding is concatenated at the beginning. The encoded class embedding is then input to a MLP Head to predict the class label after going through the Transformer encoder.

The Transformer encoder, as shown on the left part of figure 3.1, is composed of two parts. The first part is the Multi-Head Self- Attention which aggregates spacial information through Self-Attention mechanism. In Self-Attention mechanism, the patch embeddings are first linear transformed to query, key and value embeddings. Then the attention score between each patch pair is computed by the inner products of each query embedding and key embedding. With the attenion scores as weights, the new patch embedding is the weighted sum of all other patches' value embeddings:

$$\mathrm{Attention}(Q, K, V) = \mathrm{Softmax}(\frac{QK^T}{\sqrt{d}})V \tag{3.1}$$

where Q, K, V are query, key and value embeddings, $\sqrt{d}$ denotes the vector dimension of key and query embeddings. The second part is the MLP layer, also called feed-forward net, which is composed of two linear layers that aggregate the information in channel dimension.

3.2. Swin-Transformer Block

The Swin-Transformer Block follows the same procedure as described on the left part of Figure 3.1. However, in order to reduce the complexity, it first divides the image into windows and then performs the Self-Attention mechanism within each window. The comparison between the computational complexity of the window-based Multi-Head Self-Attention and the vanilla Multi-Head Self-Attention mechanism is indicated by the Formula 3.2:

$$\begin{aligned}\Omega(\mathrm{MSA}) &= 4hwC^2 + 2(hw)^2C, \\ \Omega(\mathrm{WMSA}) &= 4hwC^2 + 2M^2hwC,\end{aligned} \tag{3.2}$$

where h, w, C, M are the height, width, channel dimension, window size of the feature map respectively. Furthermore, to enable the interaction between patches in different windows, it adapts two different window partitions successively, as shown in figure 3.2.

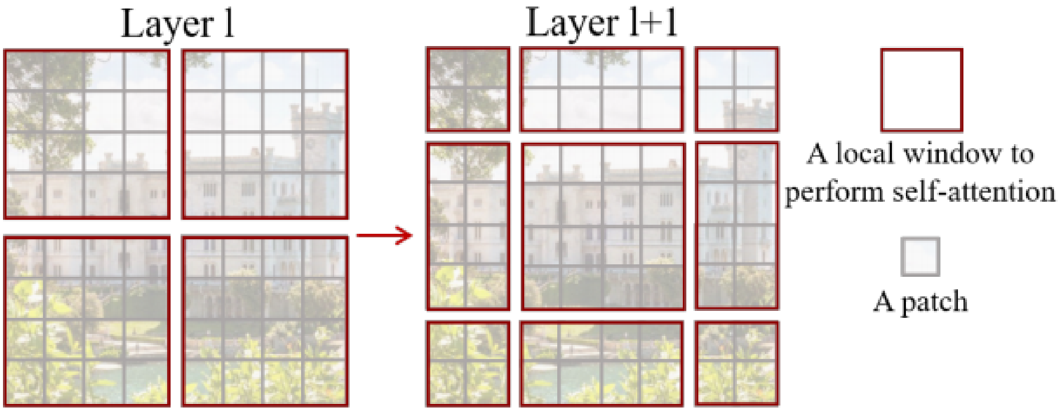


Figure 3.2: Two different window partitions for the Swin-Transformer block. The picture is adapted from the Swin-Transformer paper[14]

In order to compute the second window partition efficiently, the irregular window partition needs to be turned into regular windows by cyclic shift. The regular windows partition means that each window has the same shape so that they can be computed in parallel. Before discussing the cyclic shift, we need to remember that each patch only interacts with other patches within the same window. Thus, we have to ensure this property after converting the irregular window partition to the regular partition. The process of cyclic shift is shown in Figure 3.3, as indicated by the word 'cyclic', left-top patches A are shifted to the right-bottom while the top and left patches, C and B, are shifted to the bottom and right. In the implementation, the cyclic shift is $\lfloor \frac{M}{2} \rfloor$ units, where M is the window size. For example, the feature map in Figure 3.3, is shifted for 2 units with the window size 4. After cyclic shift, the irregular window partition is turned into four regular windows. To avoid the interactions of patches in different windows, we need to apply the masks when computing the Self-Attention in new windows. For instance, as shown in Figure 3.3, the right bottom window after cyclic shift consists of patches from top-left A, left edge B, and top edge C. They are not supposed to be the neighbors in the original irregular partition. Thus, we apply a mask to the attention matrix to manually assign the attention score between A,B,C and grey patches to a minimum value that can be ignored. After performing the Self-Attention mechanism, we reverse the window partition to the original partition by reverse cyclic shift so that the feature map is recovered.

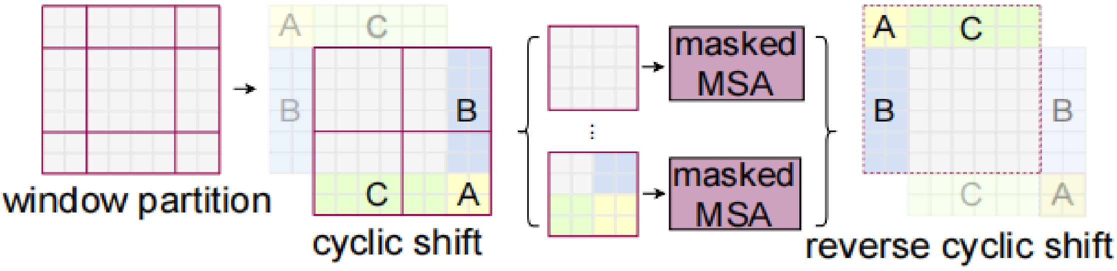


Figure 3.3: The cyclic shift process of the Swin-Transformer. The picture is adapted from the Swin-Transformer paper[14]

3.3. nnFormer

nnFormer[22] is a network architecture that combines the Swin-Transformer[14] and nnUNet[11]. The Figure 3.5 shows that overview of the nnFormer network. An encoder, a neck and a decoder together form a U-shaped architecture. The encoder consists of a patch embedding layer and four 3D Swin-Transformer blocks while the decoder is composed of three 3D Swin-Transformer block and a patch expanding layer. A sampled sub-volume of the 3D CT scan is input to the patch embedding layer which consists of two consecutive convolution blocks to extract low-level features. The feature map is then input to the 3D Swin-Transformer blocks to further extract feature maps while reducing the resolution. The output of the encoder is then used as the input to the decoder as a neck. Subsequently, the Swin-Transformer blocks in the decoder fuse the feature map from the former layer/neck and the residual connection from the encoder. As the feature map gradually recovers the resolution, it is finally passed through the patch embedding layer to perform the Trans-convolution to predict the masks.

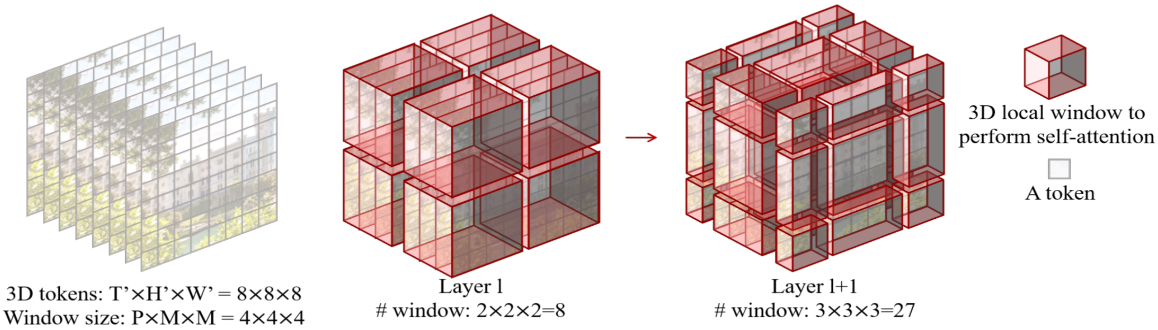

3D local window to perform self-attention
A token
3D tokens: T'×H'×W' = 8×8×8
Window size: P×M×M = 4×4×4
Layer l
# window: 2×2×2=8
Layer l+1
# window: 3×3×3=27

Figure 3.4: The window partition for 3D Swin-Transformer: the total number of windows is 8 in the regular partition, while the total number of windows is 27 in the irregular partition. The picture is adapted from the video Swin-Transformer paper[**liu2022swin**]

3D Swin-Transformer blocks as a fundamental building blocks are used in nnFormer[22] . In fact with a few lines of modification in performing the Self-Attention mechanism, a 2D Swin-Transformer block can be converted to 3D Swin-Transformer block easily. The same structure is also used in videos for human action recognition[15]. As shown in Figure 3.4, the window partition is similar to it in 2D case except for one more dimension in 3D case.

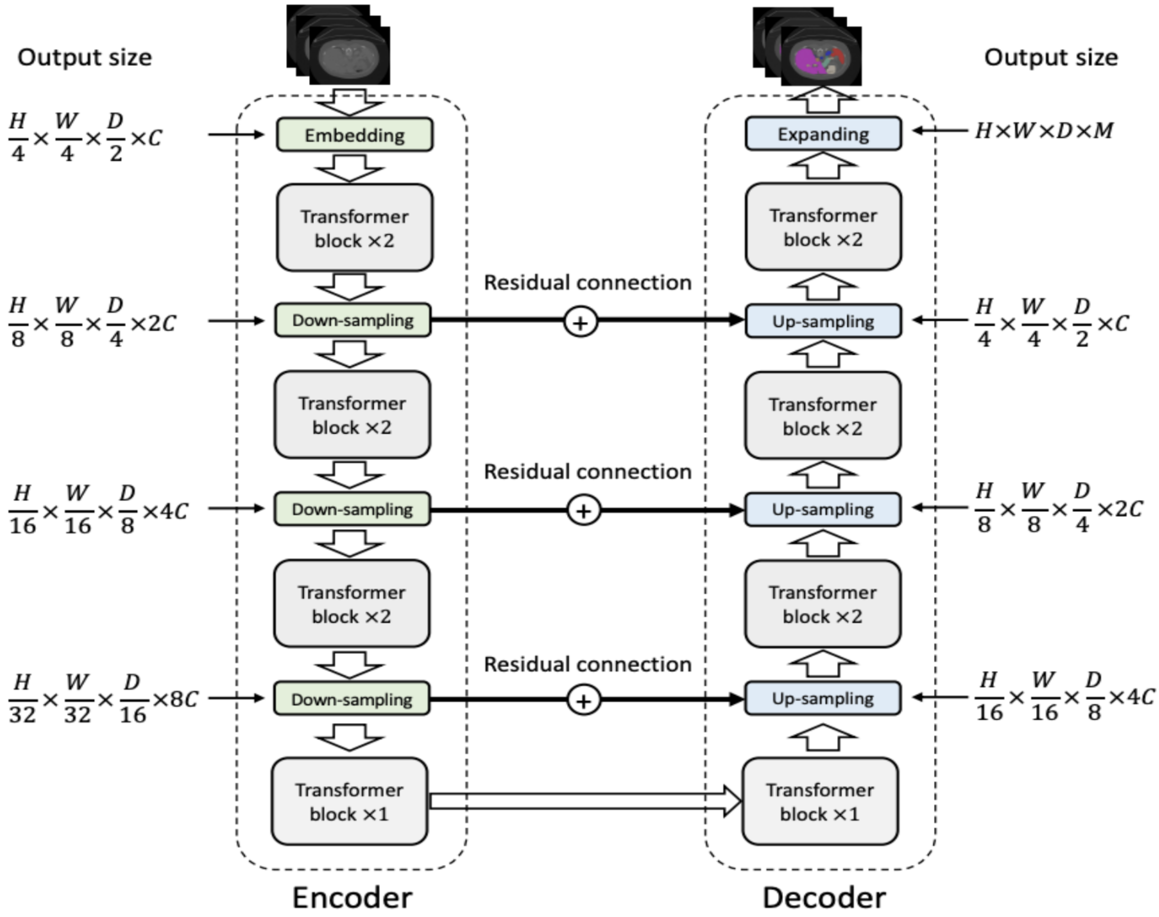

Output size
$\frac{H}{4}\times\frac{W}{4}\times\frac{D}{2}\times C$
Embedding
Transformer block ×2
$\frac{H}{8}\times\frac{W}{8}\times\frac{D}{4}\times 2C$
Down-sampling
Transformer block ×2
$\frac{H}{16}\times\frac{W}{16}\times\frac{D}{8}\times 4C$
Down-sampling
Transformer block ×2
$\frac{H}{32}\times\frac{W}{32}\times\frac{D}{16}\times 8C$
Down-sampling
Transformer block ×1
Encoder
Residual connection
Residual connection
Residual connection
Output size
$H\times W\times D\times M$
Expanding
Transformer block ×2
Up-sampling
$\frac{H}{4}\times\frac{W}{4}\times\frac{D}{2}\times C$
Transformer block ×2
Up-sampling
$\frac{H}{8}\times\frac{W}{8}\times\frac{D}{4}\times 2C$
Transformer block ×2
Up-sampling
$\frac{H}{16}\times\frac{W}{16}\times\frac{D}{8}\times 4C$
Transformer block ×1
Decoder

Figure 3.5: The overview of the nnFormer architecture. The picture is adapted from the nnFormer paper[22]

3.4. Position Encodings

The self-attention mechanism itself is permutation-invariant to the input sequences [19]. This means that the position of input image patches are not taken into account in the Self-Attention mechanism. In particular, we compute the attention score of two patches by using the key and query embeddings which both represent content information. However, it is possible that the distance of two patches matters. For example, the neighbor patches might have a higher attention value than remote patches. Therefore, in order to incorporate the position information into patch embeddings, we need to mannually introduce the position encodings to the Transformer architecture. We introduce the position encodings discussed in our project. However, there are other position encodings, such as CPT[5] which uses convolutions to extract position information, Learnable Fourier position embeddings[12]. Thus, We regard it as one of our limitations.

3.4.1. Absolute Position embedding

The absolute position embeddings can be divided into two categories: learned and fixed. The learned absolute position embeddings used in Swin-Transformer[14] and ViT[19] have the same shape as the output feature map of patch embedding. The embedding values are randomly sampled from a normal distribution with mean 0 and standard deviation 0.02. Sinusoid position embedding is the fixed absolute position embedding. We can extend original 1D sinusoid position embedding[19] in Formula

3.3 to 2D and 3D case by following the Formula 3.4 3.5.

$$\begin{aligned} \mathrm{PE(pos, 2i)} &= \mathrm{sin}(x/10000^{2i/D}) \\ \mathrm{PE(pos, 2i+1)} &= \mathrm{cos}(x/10000^{2i/D}) \end{aligned} \tag{3.3}$$

$$\begin{aligned} \mathrm{PE(x, y, 2i)} &= \mathrm{sin}(x/10000^{4i/D}) \\ \mathrm{PE(x, y, 2i+1)} &= \mathrm{cos}(x/10000^{4i/D}) \\ \mathrm{PE(x, y, 2j+D/2)} &= \mathrm{sin}(y/10000^{4j/D}) \\ \mathrm{PE(x, y, 2j+1+D/2)} &= \mathrm{cos}(y/10000^{4j/D}) \end{aligned} \tag{3.4}$$

$$\begin{aligned} \mathrm{PE(x, y, z, 2i)} &= \mathrm{sin}(x/10000^{6i/D}) \\ \mathrm{PE(x, y, z, 2i+1)} &= \mathrm{cos}(x/10000^{6i/D}) \\ \mathrm{PE(x, y, z, 2j+D/3)} &= \mathrm{sin}(y/10000^{6j/D}) \\ \mathrm{PE(x, y, z, 2j+1+D/3)} &= \mathrm{cos}(y/10000^{6j/D}) \\ \mathrm{PE(x, y, z, 2k+2D/3)} &= \mathrm{sin}(z/10000^{6k/D}) \\ \mathrm{PE(x, y, z, 2k+1+2D/3)} &= \mathrm{cos}(z/10000^{6k/D}) \end{aligned} \tag{3.5}$$

where x,y,z denote the corresponding coordinates on the feature map, D denotes the dimension of the position embedding, i,j,k denote the value location within embedding dimension(channel dimension).

3.4.2. Relative Position Bias

As shown in Formula 3.6, the relative position bias[14] is added when computing the attention matrix in each Swin-Transformer block. Each Swin-Transformer layer constructs a look-up table to store all possible relative position biases within a window. A layer with window size M has total number of $(2M-1) \times (2M-1) \times (2M-1)$ relative positions. The bias is later added to the attention score according to the relative position between query and key embeddings.

$$\mathrm{Attention(Q, K, V)} = \mathrm{Softmax}(\frac{\mathrm{QK^T}}{\sqrt{\mathrm{d}}} + \mathrm{B})\mathrm{V} \tag{3.6}$$

3.5. Data

In this section, we briefly introduce the prostate CT data set we used in our project. We use prostate CT data containing annotations of four organs: bladder, prostate, rectum, and seminal vesicles. The data is collected from three institutes, c.f. Haukeland Medical Center of Norway (HMC), Leiden University Medical Center in the Netherlands (LUMC) and Erasmus Medical Center in the Netherlands (EMC), containing 179, 475 and 56 CT scans, respectively. EMC is used as the test data set, while HMC and LUMC are used as the training datasets. Due to differences in clinical protocols for CT scan acquisition, the EMC dataset has larger volumes of the prostate and bladder, which makes it a challenging test dataset.

As mentioned above, due to different clinical protocols for acquisition, prostate CT scans from different institutions have differences. The differences of three datasets in organ voxel numbers are shown in figure 3.6. Patients are required to drink water before taking CTs by Erasmus Medical Center so the EMC CT scans have a larger number of voxels in prostate and bladder. Apart from statistics, the 3D visual differences of three datasets are shown in Figure 3.7.

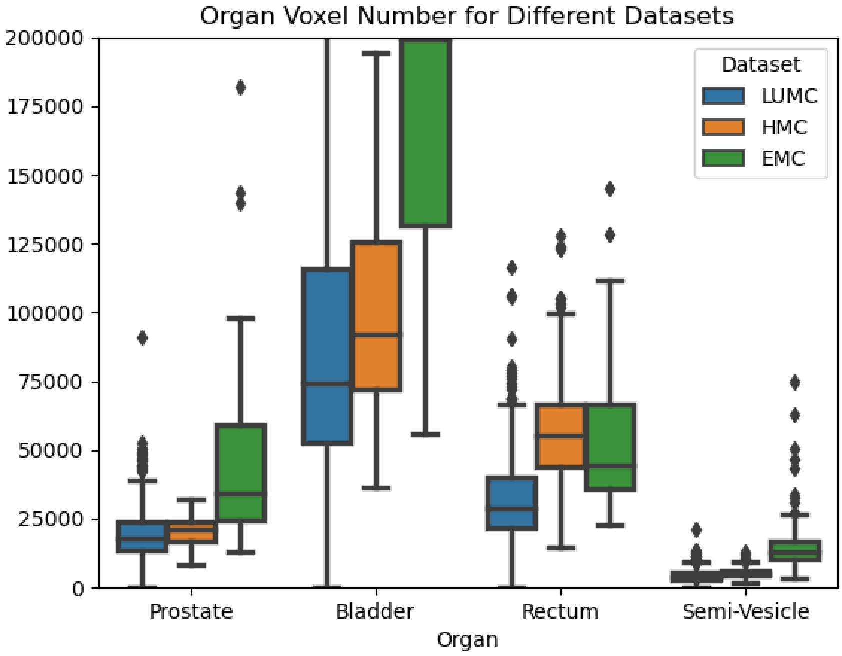


Figure 3.6: The number of voxels in each organ in three different datasets.

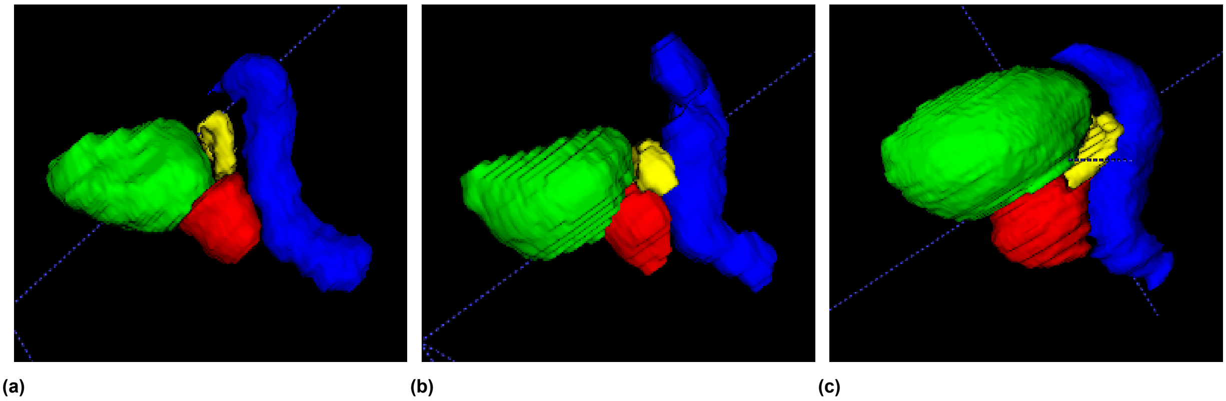


Figure 3.7: Visual examples: (a) from HMC dataset, (b) from LUMC dataset, (c)from EMC dataset. The green part is bladder; the yellow part is seminal vesicle; the red part is prostate; the blue part is rectum.

References

[1] Hangbo Bao, Li Dong, and Furu Wei. “Beit: Bert pre-training of image transformers”. In: *arXiv preprint arXiv:2106.08254* (2021).

[2] Tom Brown et al. “Language models are few-shot learners”. In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

[3] Hu Cao et al. “Swin-unet: Unet-like pure transformer for medical image segmentation”. In: *arXiv preprint arXiv:2105.05537* (2021).

[4] Zhe Chen et al. “Vision Transformer Adapter for Dense Predictions”. In: *arXiv preprint arXiv:2205.08534* (2022).

[5] Xiangxiang Chu et al. “Conditional positional encodings for vision transformers”. In: *arXiv preprint arXiv:2102.10882* (2021).

[6] Zihang Dai et al. “Coatnet: Marrying convolution and attention for all data sizes”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 3965–3977.

[7] Jacob Devlin et al. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805* (2018).

[8] Alexey Dosovitskiy et al. “An image is worth 16x16 words: Transformers for image recognition at scale”. In: *arXiv preprint arXiv:2010.11929* (2020).

[9] Ali Hatamizadeh et al. “Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images”. In: *International MICCAI Brainlesion Workshop*. Springer. 2022, pp. 272–284.

[10] Ali Hatamizadeh et al. “UNETR: Transformers for 3d medical image segmentation”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2022.

[11] Fabian Isensee et al. “nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation”. In: *Nature methods* (2021).

[12] Yang Li et al. “Learnable fourier features for multi-dimensional spatial positional encoding”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 15816–15829.

[13] Ze Liu et al. “Swin transformer v2: Scaling up capacity and resolution”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 12009–12019.

[14] Ze Liu et al. “Swin transformer: Hierarchical vision transformer using shifted windows”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021.

[15] Ze Liu et al. “Video swin transformer”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 3202–3211.

[16] Zhuang Liu et al. “A convnet for the 2020s”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

[17] Christos Matsoukas et al. “Is it time to replace cnns with transformers for medical images?” In: *arXiv preprint arXiv:2108.09038* (2021).

[18] Ikboljon Sobirov et al. “Automatic Segmentation of Head and Neck Tumor: How Powerful Transformers Are?” In: *Medical Imaging with Deep Learning*. 2022.

[19] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017).

[20] Haiping Wu et al. “Cvt: Introducing convolutions to vision transformers”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 22–31.

[21] Hao Zhang et al. “Dino: Detr with improved denoising anchor boxes for end-to-end object detection”. In: *arXiv preprint arXiv:2203.03605* (2022).

[22] Hong-Yu Zhou et al. “nnformer: Interleaved transformer for volumetric segmentation”. In: *arXiv preprint arXiv:2109.03201* (2021).