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

Yicong Tan[a], Prerak Mody[b], Viktor van der Valk[b], Marius Staring[b,c], and Jan van Gemert[a]

[a]Pattern Recognition Lab, TU Delft, Delft, The Netherlands
[b]Department of Radiology, Leiden University Medical Center, Leiden, The Netherlands
[c]Department of Radiation Oncology, Leiden University Medical Center, Leiden, The Netherlands

ABSTRACT

Literature on medical imaging segmentation claims that hybrid UNet models containing both Transformer and convolutional blocks perform better than purely convolutional UNet models. This recently touted success of hybrid Transformers warrants an investigation into which of its components contribute to its performance. Also, previous work has a limitation of analysis only at fixed dataset scales as well as unfair comparisons with other models where parameter counts are not equivalent. Here, we investigate the performance of a hybrid Transformer network i.e. the nnFormer for organ segmentation in prostate CT scans. We do this in context of replacing its various components and by constructing learning curves by plotting model performance at different dataset scales. To compare with literature, the first experiment replaces all the shifted-window(swin) Transformer blocks of the nnFormer with convolutions. Results show that the convolution prevails as the data scale increases. In the second experiment, to reduce complexity, the self-attention mechanism within the swin-Transformer block is replaced with an similar albeit simpler spatial mixing operation i.e. max-pooling. We observe improved performance for max-pooling in smaller dataset scales, indicating that the window-based Transformer may not be the best choice in both small and larger dataset 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 hybrid swin-Transformers deficiency in capturing positional information given our dataset at its various scales. Through this work, we hope to motivate the community to use learning curves under fair experimental settings to evaluate the efficacy of newer architectures like Transformers for their medical imaging tasks. Code is available on https://github.com/prerakmody/window-transformer-prostate-segmentation.

Keywords: Radiotherapy, Segmentation, 3D Swin-Transformer, Convolution, Pooling, Positional Encoding Learning curves

1. INTRODUCTION

Transformer[1] networks are popular in both natural language processing[2,3] and vision domains.[4–10] In the vision domain, Vision Transformer[4] set the foundation by creating a general structure for applying the Transformer blocks in image classification tasks. Others have broadened the use of the Transformer by addressing the limitations in a Vanilla Vision Transformer by either reducing the complexity in the self-attention mechanism[11–14] or by introducing inductive biases from convolutions to improve performance.[8,9,14–17] Encouraged by these successes in the vision domain, researchers are attempting to apply the Transformer to medical images.

Cancer treatment via radiotherapy requires the segmentation of tumors and organs-at-risk (OAR) on diagnostic scans like CT. Convolution-based UNet architectures[18] 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.[19–23] In particular, do they offer an alternative for

Further author information: (Send correspondence to P.P.M.)
P.P.M.: E-mail: P.P.Mody@lumc.nl

the locality bias of convolutions which plays an important role in medical image segmentation where one needs to precisely contour the borders of a region of interest (i.e. organs and tumors).

Specifically for 3D medical image segmentation, we investigate a popular shifted-window(swin)-based hybrid Transformer i.e. nnFormer[19] which contains convolutional and transformer blocks in an interleaved manner. We chose this model since it uses the shifted-windows[14] concept in its transformer block. They were originally proposed since the vanilla Transformer suffered from a computational complexity quadratic to the image size, which is a major deterrent to processing 3D medical images. Also, unlike other hybrid-UNet models like UNETR[20] which restrict the transformer blocks to the encoder, the nnFormer performs alternate convolutions and transformer operations, hence taking advantages of both modules. However, due to the scarcity of medical data, one of the issues with the proposals of existing hybrid-Transformers[19,20] is that their training datasets could potentially be insufficient for the task at hand, leading to overfitting. Moreover, previous work does not provide any learning curves for their datasets to investigate the role of dataset scales. Learning curves help identify at what point additional data provides diminishing returns. Our work remedies this by analyzing 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 of previous work is that the Transformer and convolutions are compared in a different general neural structure without ensuring equivalent parameter count.[19,20,22] For example, the nnFormer model with 150M parameters is compared to a nnUNet with 30M parameters.[19] Inspired by ConvNeXt[24] 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 hybrid swin-Transformers do not offer any performance gains over the comparison models 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 the UNet, which is the widely-used architecture in medical image segmentation. UNet,[25] first applied in 2D slices and then extended to 3D CT scans and MRI images,[26] 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[27] and DenseNet[28] stimulated the development of ResUNet,[29] ResUNet++,[30] Multi-ResUNet[31] and DenseUNet.[32] Additionally, the aggregation of the output from deep and shallow layers inspired the UNet++.[33,34] Besides, the combination of attention mechanisms and UNet resulted in the Attention-UNet,[35] Attention-UNet++,[36] MA-UNet,[37] SCAU-Net,[38] and AA-UNet.[39] All these methods successfully incorporated novel ideas into UNet architecture and attained improvement, providing the intuition for importing Transformer blocks into UNet.

The success of the 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,[40] TransUnet,[41] MCTrans[42] and TransClaw U-Net.[43] In the meanwhile, transferring the architecture directly from the vision domain has become a trend, for example, Swin-Transformer[14] to Swin-UNet[44] and LeVit[45] to LeVit-UNet.[46] MedT[47] and MBT-Net[48] adopted the Axial-Attention[11] block to reduce the computational complexity while taking advantage of the Self-Attention mechanism. Despite the improvement in 2D medical segmentation brought by Transformer blocks, there is a large gap between these 2D models and the 3D convolutional benchmark nnUNet.[18]

Researchers have also recently applied the Transformer to 3D medical segmentation. However, this process is still at the initial stage. TransBTS[49] used the Transformer block to fuse the feature maps from 3D ConvNets. Besides, UNETR[20] replaced the convolution blocks with Transformer blocks in the 3D-UNet encoder. On top of that, Swin-UNETR[50] changed the vanilla Transformer blocks with swin-Transformer blocks. Similarly, nnFormer[19] replaced convolutions with swin-Transformer blocks in both the encoder and the decoder and incorporated the model in the frame of the nnUNet.[18] In addition, D-Former[51] was inspired by the dilated convolution and restricted the Self-Attention in a dilated block to reduce the complexity and enlarge the receptive field. In this paper, we also built our experiment based on the nnUNet[18] that provides a general framework to handle

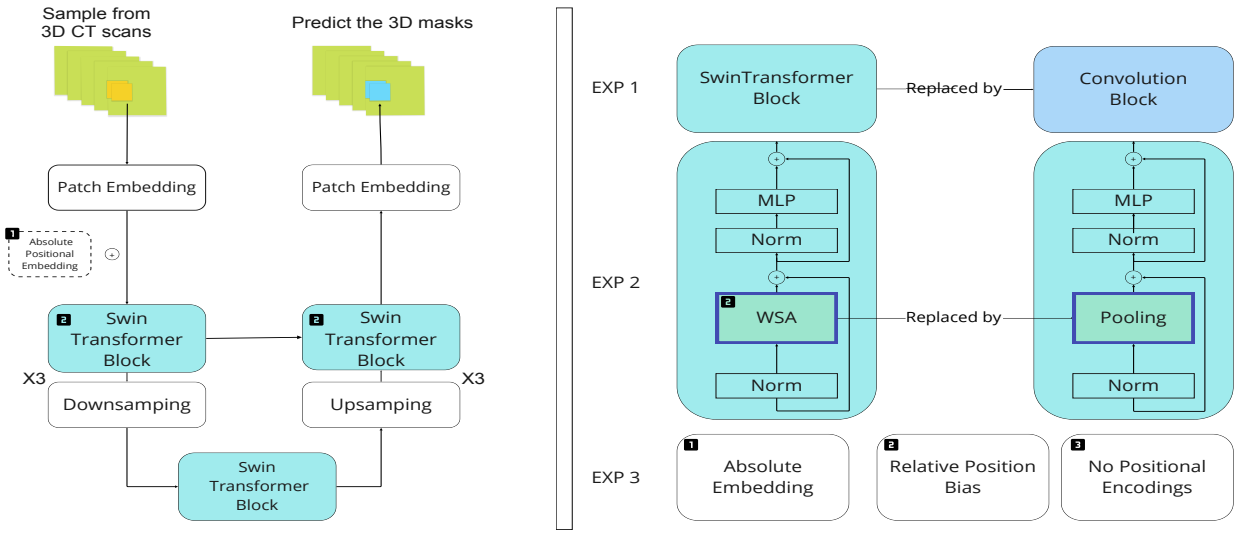


Figure 1: This figure shows the network architecture of the nnFormer[19] along with modifications made for our three experiments. The left part, the yellow tiles denote the sampled sub-volumes from the 3D CT scans while the blue tiles denote the predicted masks for segmentation. On the right, we show the components of the nnFormer and their replacements, which we evaluate in experiments EXP1, EXP2, and EXP3. In EXP1, we replace the Swin-tTransformer block with the convolution block. In EXP2 the orange blocks show that we only replace the window-based self-attention with pooling. In EXP3 we use labels with black background to show the place to insert these position encodings: the absolute position embeddings are added once to each position on the feature map after patch embedding and the relative bias is added in computing self-attention matrix in each Swin-Transformer block.

arbitrary medical image 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.[21,23] The method introduced by Matsoukas[23] was restricted to 2D segmentation. 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[24] 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,[52] MLP-Mixer[53,54] and Conv-Mixer.[55] They split the Transformer block into two parts: the Self-Attention corresponds 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 Network Architecture

As shown in Figure 11, the hybrid-Transformer network in evaluation is the nnFormer,[19] which employs Swin (shifted-window)-Transformer blocks[14,44] interleaved with convolutions 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. As shown in the left part of Figure 11, we have three experimental settings to examine the different components of the hybrid Swin-Transformer network. First, two Swin-Transformer blocks in one layer are replaced by two 3 x 3 x 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.

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

Literature on medical image segmentation has shown superior performance of window-based Transformers over convolutions.[19,20] We test this notion by replacing the window-based Transformer blocks with a sequence of two convolutions. 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.

3.1.2 Method 2: Replacing the Self-Attention with Pooling

In the spirit of further analyzing components of the Transformer block and inspired by the MetaFormer[52] to reduce computational complexity, we replace the attention mechanism with a much simpler spatial feature mixing operation, i.e pooling. 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.

3.1.3 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 to a model without any positional encodings. The first is absolute positional embedding that is added to the feature map after the convolutional patch-embedding as shown in Figure 1. 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 the original 1D sinusoid[1] positional embedding to 3D case. 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.[19]

4. EXPERIMENTS AND RESULTS

4.1 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. An argument can be made that datasets of human body scans do not need to be as large as natural image datasets. Although there exists differences across patients in the shape and size of individual organs, they are still very similar in context of spatial arrangement of organs. Thus, we consider our dataset to be sufficient to analyze the learning curves of the hybrid-Transformers and its variations.

To test a different modality, we also conduct experiments on two open datasets of prostate MR c.f. ProstateX[56] and PROMISE12[57] containing 66 and 50 scans respectively. This public availability of this dataset shall also allow the community to reproduce some of our experiments.

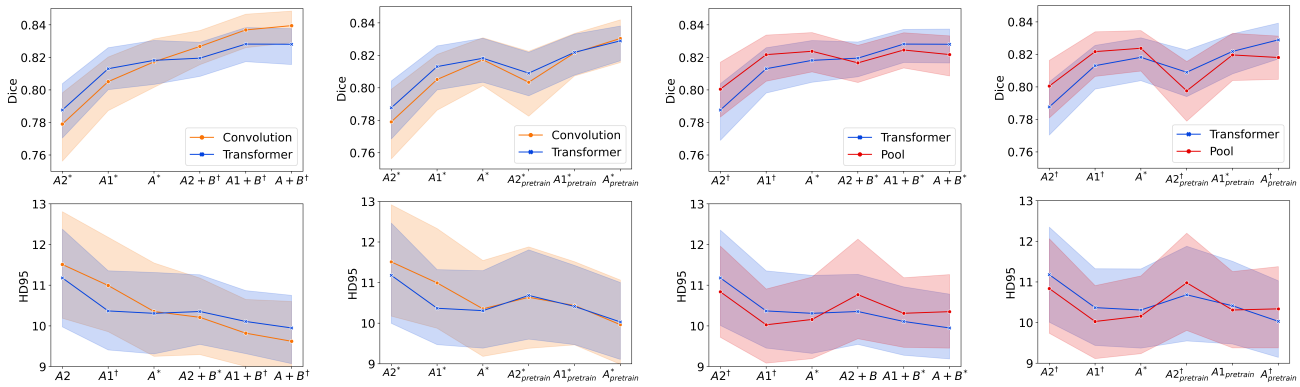


Figure 2: Here we see line plots showing the mean and 95$^{\text{th}}$ percentile confidence interval of Dice and HD95 for (a)(b) experiment 1 wherein we replace the Swin-transformer block with the convolutional block and for (c)(d) experiment 2 where the attention mechanism within the Swin-transformer block is replaced with pooling: The x-axis denotes the various training data scales with clinics A and B while A2, A1 denote the first and second halves 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.

4.2 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 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$^{\text{th}}$ percentile Hausdorff Distance (HD95) averaged over all scans of the test dataset. In addition, we adopt Wilcoxon signed rank test with p value less than 0.5 to reveal the statistical significance between the results of two compared models.

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, since this is the count of parameters in the original nnFormer work.[19] The CT scans are first resampled to the median spacing of each dataset and then randomly sampled patches of size (128,128,64) are input to the network. Models were trained with Pytorch 11.3 on a single Nvidia RTX6000 (24GB memory). In addition to the line plots in Figure 2 and 3, detailed experimental results are reported in the Appendix.

4.3 Experiment 1: Replacing Swin-Transformer block with convolution block

Surprisingly, the results in Figure 2 (a) and contours in Figure 6 (a)(b) show that the Transformer performs better on lower data scales and the convolution gradually surpasses it with the increase of data. Perhaps, 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 neighborhood of a voxel (*disconnected components of seminal vesicle in Figure 6 (a)*). 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 dataset regimes. The higher performance of convolutions in our larger data scales may imply that the Transformer needs even more data than available in our dataset to understand the local structure of the semantic content within the data (*jagged nature of bladder in Figure 6 (a)*). Moreover, we observed that the stability of transformers is quite poor, with sudden drops in performance during training.

In the pretraining context, Figure 2 (b) suggest that both the Transformer and the convolution benefit from pretraining (e.g. $A2$ vs $A2_{pretrain}$, which is obvious due to more exposure to the data. Visual results also indicate the same for e.g. the lower sections of prostate (red) in Figure 6 (e) vs Figure 6 (f). However, at the same dataset scale for e.g. $A2$+B vs $A2_{pretrain}$, the gaps between the Transformer and convolution are eliminated after pretraining primarily due to lower score of convolutions. Since pretraining essentially breaks supervision into two phases, the slightly lower performance due to pretraining makes sense. The visual results in 6 (e),(f) indicate the same where the non-smooth boundaries of the prostate(red) are very different compared to the ground truth for both transformers and convolutions.

We conduct additional experiments to verify the robustness in the trends of our learning curve results by halving the parameters of both the Transformer and convolution models to 75M. Figure 5 (b) shows similar results to that of Figure 2 (a) where we see that the Transformer performs better on lower dataset scales and the convolution gradually surpasses it with increased data.

4.4 Experiment 2: Replacing the Self-Attention with Pooling

In line with our expectations, Figure 2 (c) suggest that max-pooling outperforms in smaller dataset scales compared to the window-based self-attention mechanism, while the latter surpasses it with the increase of dataset scale. Thus, both convolution and window-based Transformer overfit under small dataset scales in comparison to pooling. These results indicate that the simplicity of pooling may be essential to high performance in small dataset regimes. Visually, we can confirm this by observing the contours of the bladder (green) in Figure 6 (c) and (d) for the smallest and largest dataset scale respectively.

Apart from that, Figure 2 (d) show that the Transformer model benefits from pretraining, while the pooling operation fails to elevate the performance even with pretraining (e.g. $A2$ vs $A2_{pretrain}$). This can be seen by the poorer contour performance of bladder (green) in the 6 (h) for the pooling operation when compared to 6 (g). This might further indicate that a larger dataset scale compared to our clinic A is desired for our Transformer model.

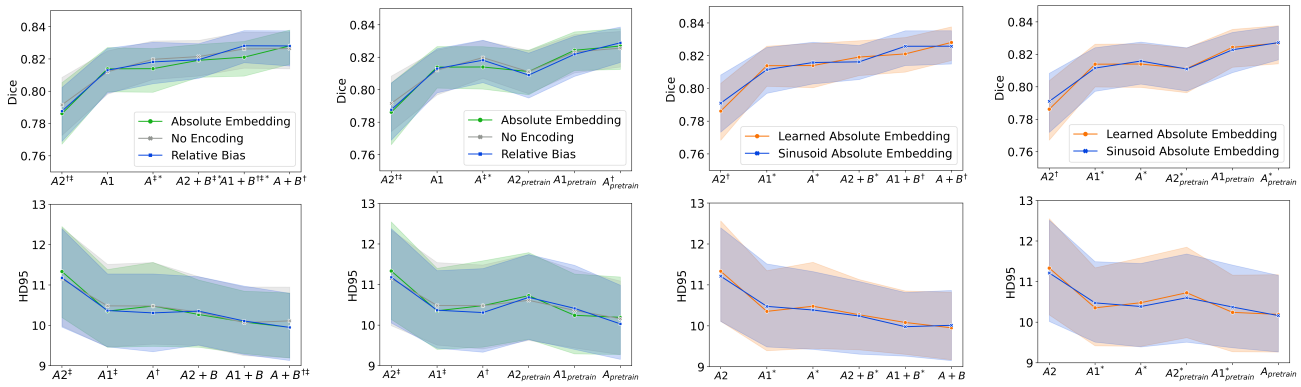
In addition, we have compared the average-pooling and max-pooling in the smallest three data scales c.f. A1, A2 and A. Figure 5(d) indicates that max-pooling has an advantage over the average-pooling. A possible explanation for this is that the max-pooling extracts the important local features (e.g. edges) and highlights them while average-pooling smooths them and hence attenuates the useful signals produced by the convolutional filter preceding it.

4.5 Experiment 3: Evaluating Positional Encoding

Figure 3 (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 image segmentation. Contrary to Transformers, both convolutions and pooling have some form of positional inductive bias (i.e. locality and neighborhood structure). This could be one reason that the window-based Transformer does not offer any benefits over the comparison models in both our smaller and larger dataset scales, which is contrary to claims in literature.[19]

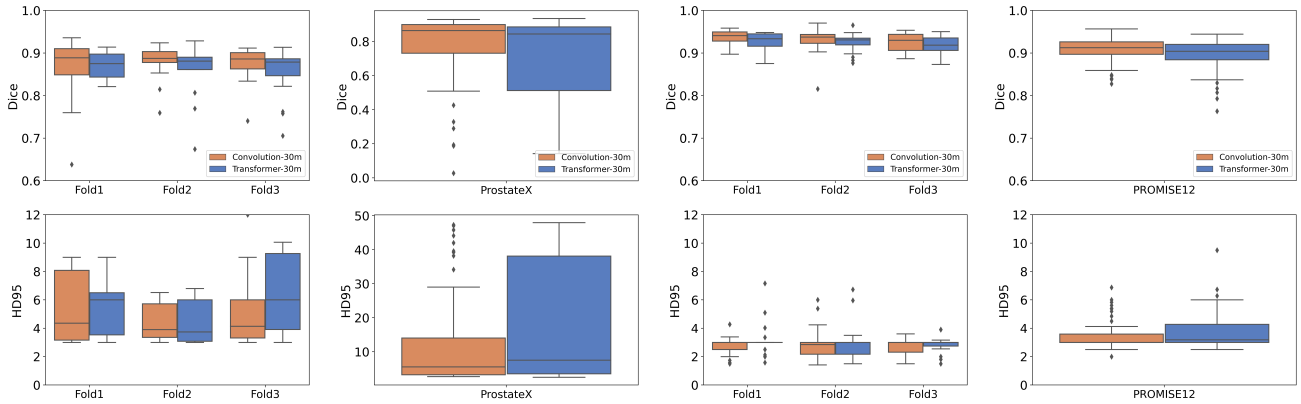
4.6 Ablation Experiments

Given the wide acceptance of the nnUNet[18] model, we do a fair comparison of the nnFormer and its convolutional version with it at its parameter count of 30M. Figure 5 (a) shows that nnUNet performs slightly worse in smaller dataset scales, but being a fully convolutional model, it eventually surpasses it in larger dataset scales. This is an important observation as literature often treats nnUNet as a baseline model but does not equalize parameter count. Independent of the other experiments and in the spirit of the title of this work, we also investigate another component of transformers c.f. layer normalization. Figure 5 (c) shows that removing the layer normalization from the Transformer model decreases the performance, and this effect is exacerbated in large dataset scales. This is an expected result as it has been previously shown that a neural network requires normalization layers as they help conserve the learning signal being backpropagated in deep neural networks.



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

Figure 3: Experiment 3: 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. 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.

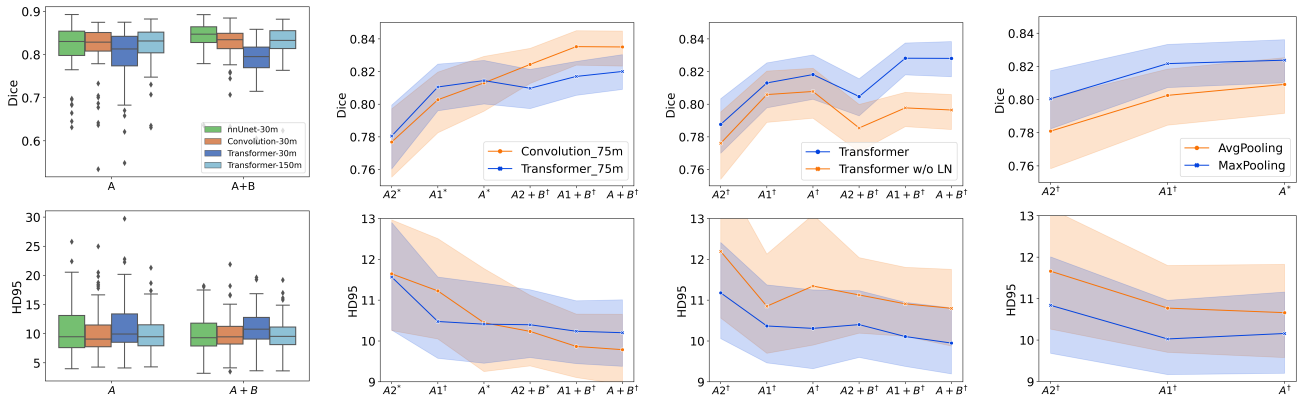


(a) (b) (c) (d)

Figure 4: Here we see boxplots comparing the nnFormer and its convolutional version pertaining to experiments on the open datasets – ProstateX and PROMISE12. In (a), (c), three-fold cross-validation is conducted on PROMISE12 and ProstateX respectively and the validation fold results are shown. In (b) the models are trained on all samples of ProstateX and then tested on all samples of PROMISE12, while in (d) the reverse is done. The x-axis labels on (b) and (d) show the training dataset.

4.7 Open Dataset

In addition to the private dataset, we compare the nnFormer and its convolutional version with 30M parameters on the public datasets – PROMISE12[57] and ProstateX.[56] We conduct three-fold cross-validation on both datasets separately and also conduct another experiment where one is used a train and the other as a test dataset. The results on Figure 4 (a), (b), (c), (d) show that convolutional version with 30M parameters performs slightly better on both datasets in our two experiment settings. We notice much poorer performance when trained on ProstateX and tested on PROMISE12, since many of its scans were taken with the presence of rectal coil, a scanning protocol not present in ProstateX. Thus, the results in 4 show that the Swin-Transformer block does not provide any gains in our experiments.



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

Figure 5: Ablation 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.

5. DISCUSSION AND CONCLUSION

This study evaluates different components of a shifted-window(Swin)-based hybrid Transformer network, i.e. nnFormer to understand their role in its performance. Unlike previous work that uses hybrid networks containing both convolution and Transformer blocks,[19,20,22,51] we maintain a constant parameter count across models and also create learning curves by analyzing the effect of the components under different dataset scales and pretraining context. Our first experiments results show that the Swin-Transformer blocks do not offer any advantages over their convolutional counterparts in the larger dataset scales of this work. We believe this behaviour of Swin-Transformer blocks could be explained by their lack of understanding of positional information of voxels in our dataset scales. Our second experiments results suggest that it may always be better to choose simpler operations like pooling in small dataset regimes. Finally, in our third experiment, the comparable performance of models with and without positional encodings further supports our claim about the Swin-Transformer blocks inability to understand positional information given our dataset scales. Thus, we conclude that for our dataset, the Swin-Transformer block is not the best choice in both small and larger dataset scales. Please note that our largest data scale may not be sufficient for Transformers-based networks, which are well-known to be data hungry. However, perhaps medical image segmentation of organs does not require large datasets as has been shown for natural images. We hypothesize that the homogeneity in the 3D spatial positioning of regions of interest in human body organs, may place less stringent requirements, in context of dataset sizes. Finally, future work could use our learning curve approach under fair experimental settings to understand other network architecture for medical image segmentation. Additionally, it may be worth exploring self-supervised methods, as they could potentially benefit the data hungry Transformer.

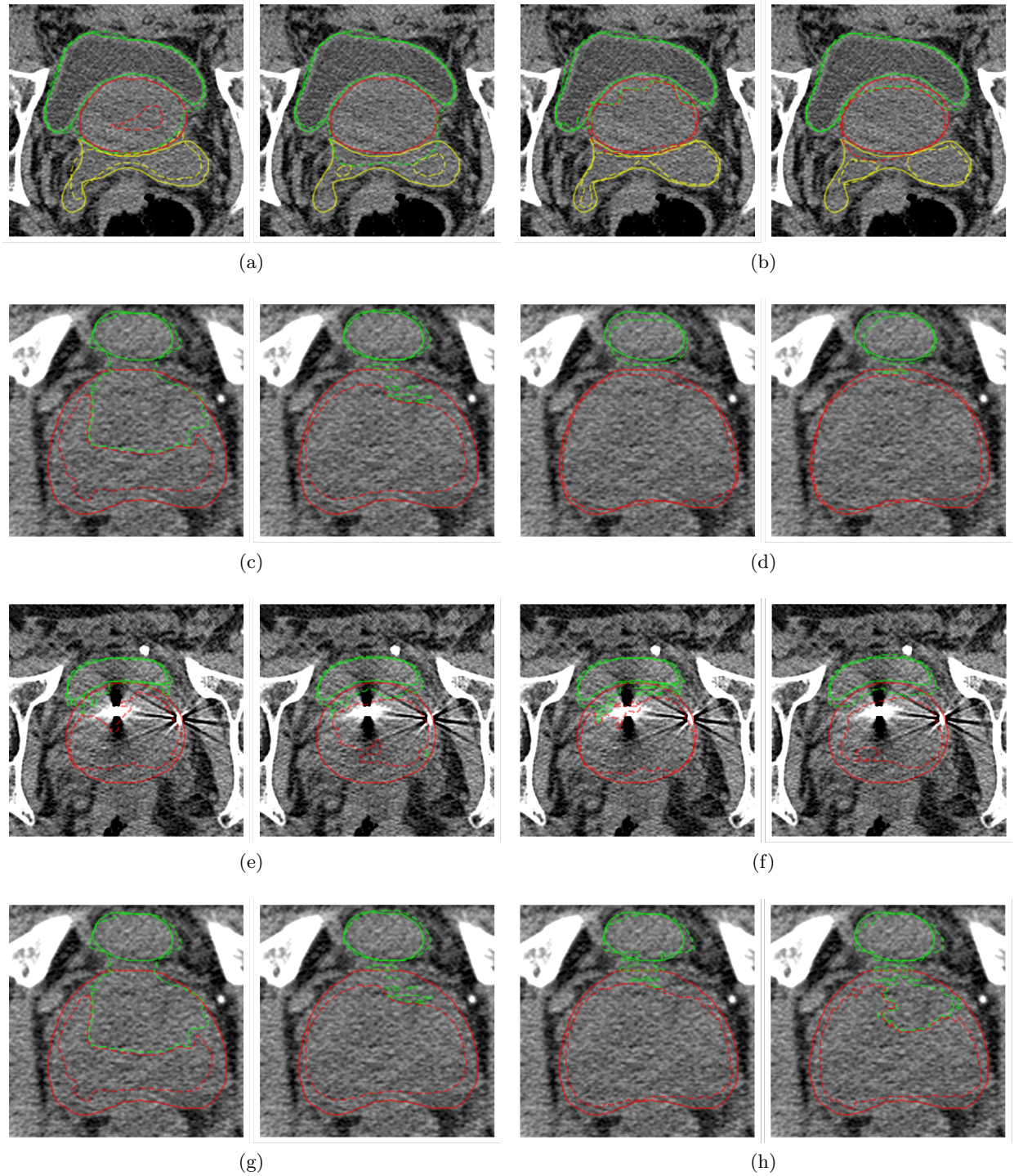


Figure 6: CT scans showing the prediction (dotted line) and ground truth (solid line) for the prostate (red), bladder (green) and seminal vesicle (yellow). The left image of each image pair is the transformer's prediction. Here a) and b) compare transformers with convolutions in the smallest and largest dataset scale respectively. Similarly, c) and d) compare transformers with pooling in the smallest and largest dataset scale, respectively. e) and f) compare the regular and pretrained models for transformers and convolutions in the A2 dataset scale respectively, while g) and h) do the same for transformers and pooling.

REFERENCES

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I., "Attention is all you need," *Advances in neural information processing systems* **30** (2017). 0, 3

[2] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K., "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805* (2018). 0

[3] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al., "Language models are few-shot learners," *Advances in neural information processing systems* **33**, 1877–1901 (2020). 0

[4] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929* (2020). 0

[5] Bao, H., Dong, L., and Wei, F., "Beit: Bert pre-training of image transformers," *arXiv preprint arXiv:2106.08254* (2021). 0

[6] Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., et al., "Swin transformer v2: Scaling up capacity and resolution," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*], 12009–12019 (2022). 0

[7] Chen, Z., Duan, Y., Wang, W., He, J., Lu, T., Dai, J., and Qiao, Y., "Vision transformer adapter for dense predictions," *arXiv preprint arXiv:2205.08534* (2022). 0

[8] Dai, Z., Liu, H., Le, Q. V., and Tan, M., "Coatnet: Marrying convolution and attention for all data sizes," *Advances in Neural Information Processing Systems* **34**, 3965–3977 (2021). 0

[9] Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., and Zhang, L., "Cvt: Introducing convolutions to vision transformers," in [*Proceedings of the IEEE/CVF International Conference on Computer Vision*], 22–31 (2021). 0

[10] Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L. M., and Shum, H.-Y., "Dino: Detr with improved denoising anchor boxes for end-to-end object detection," *arXiv preprint arXiv:2203.03605* (2022). 0

[11] Ho, J., Kalchbrenner, N., Weissenborn, D., and Salimans, T., "Axial attention in multidimensional transformers," *arXiv preprint arXiv:1912.12180* (2019). 0, 1

[12] Huang, Z., Ben, Y., Luo, G., Cheng, P., Yu, G., and Fu, B., "Shuffle transformer: Rethinking spatial shuffle for vision transformer," *arXiv preprint arXiv:2106.03650* (2021). 0

[13] Chen, C.-F., Panda, R., and Fan, Q., "Regionvit: Regional-to-local attention for vision transformers," *arXiv preprint arXiv:2106.02689* (2021). 0

[14] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B., "Swin transformer: Hierarchical vision transformer using shifted windows," in [*Proceedings of the IEEE/CVF International Conference on Computer Vision*], (2021). 0, 1, 2

[15] Xu, Y., Zhang, Q., Zhang, J., and Tao, D., "Vitae: Vision transformer advanced by exploring intrinsic inductive bias," *Advances in Neural Information Processing Systems* **34**, 28522–28535 (2021). 0

[16] Liu, Y., Sun, G., Qiu, Y., Zhang, L., Chhatkuli, A., and Van Gool, L., "Transformer in convolutional neural networks," *arXiv preprint arXiv:2106.03180* (2021). 0

[17] Chu, X., Tian, Z., Zhang, B., Wang, X., Wei, X., Xia, H., and Shen, C., "Conditional positional encodings for vision transformers," *arXiv preprint arXiv:2102.10882* (2021). 0

[18] Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., and Maier-Hein, K. H., "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods* (2021). 0, 1, 5

[19] Zhou, H.-Y., Guo, J., Zhang, Y., Yu, L., Wang, L., and Yu, Y., "nnformer: Interleaved transformer for volumetric segmentation," *arXiv preprint arXiv:2109.03201* (2021). 0, 1, 2, 3, 4, 5, 7

[20] Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H. R., and Xu, D., "UNETR: Transformers for 3d medical image segmentation," in [*Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*], (2022). 0, 1, 3, 7

[21] Sobirov, I., Nazarov, O., Alasmawi, H., and Yaqub, M., "Automatic segmentation of head and neck tumor: How powerful transformers are?," in [*Medical Imaging with Deep Learning*], (2022). 0, 2

[22] Zhang, Y., Liu, H., and Hu, Q., "Transfuse: Fusing transformers and cnns for medical image segmentation," in [*International Conference on Medical Image Computing and Computer-Assisted Intervention*], 14–24, Springer (2021). 0, 1, 7

[23] Matsoukas, C., Haslum, J. F., Söderberg, M., and Smith, K., "Is it time to replace cnns with transformers for medical images?," *arXiv preprint arXiv:2108.09038* (2021). 0, 2

[24] Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S., "A convnet for the 2020s," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*], (2022). 1, 2

[25] Ronneberger, O., Fischer, P., and Brox, T., "U-net: Convolutional networks for biomedical image segmentation," in [*International Conference on Medical image computing and computer-assisted intervention*], 234–241, Springer (2015). 1

[26] Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., and Ronneberger, O., "3d u-net: learning dense volumetric segmentation from sparse annotation," in [*International conference on medical image computing and computer-assisted intervention*], 424–432, Springer (2016). 1

[27] He, K., Zhang, X., Ren, S., and Sun, J., "Deep residual learning for image recognition," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 770–778 (2016). 1

[28] Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q., "Densely connected convolutional networks," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 4700–4708 (2017). 1

[29] Xiao, X., Lian, S., Luo, Z., and Li, S., "Weighted res-unet for high-quality retina vessel segmentation," in [*2018 9th international conference on information technology in medicine and education (ITME)*], 327–331, IEEE (2018). 1

[30] Jha, D., Smedsrud, P. H., Riegler, M. A., Johansen, D., De Lange, T., Halvorsen, P., and Johansen, H. D., "Resunet++: An advanced architecture for medical image segmentation," in [*2019 IEEE International Symposium on Multimedia (ISM)*], 225–2255, IEEE (2019). 1

[31] Ibtehaz, N. and Rahman, M. S., "Multiresunet: Rethinking the u-net architecture for multimodal biomedical image segmentation," *Neural networks* **121**, 74–87 (2020). 1

[32] Guan, S., Khan, A. A., Sikdar, S., and Chitnis, P. V., "Fully dense unet for 2-d sparse photoacoustic tomography artifact removal," *IEEE journal of biomedical and health informatics* **24**(2), 568–576 (2019). 1

[33] Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., and Liang, J., "Unet++: A nested u-net architecture for medical image segmentation," in [*Deep learning in medical image analysis and multimodal learning for clinical decision support*], 3–11, Springer (2018). 1

[34] Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., and Liang, J., "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE transactions on medical imaging* **39**(6), 1856–1867 (2019). 1

[35] Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N. Y., Kainz, B., et al., "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999* (2018). 1

[36] Li, C., Tan, Y., Chen, W., Luo, X., Gao, Y., Jia, X., and Wang, Z., "Attention unet++: A nested attention-aware u-net for liver ct image segmentation," in [*2020 IEEE International Conference on Image Processing (ICIP)*], 345–349, IEEE (2020). 1

[37] Cai, Y. and Wang, Y., "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)*], **12167**, 205–211, SPIE (2022). 1

[38] Zhao, P., Zhang, J., Fang, W., and Deng, S., "Scau-net: spatial-channel attention u-net for gland segmentation," *Frontiers in Bioengineering and Biotechnology* **8**, 670 (2020). 1

[39] Rajamani, K. T., Rani, P., Siebert, H., ElagiriRamalingam, R., and Heinrich, M. P., "Attention-augmented u-net (aa-u-net) for semantic segmentation," *Signal, image and video processing* , 1–9 (2022). 1

[40] Gao, Y., Zhou, M., and Metaxas, D. N., "Utnet: a hybrid transformer architecture for medical image segmentation," in [*International Conference on Medical Image Computing and Computer-Assisted Intervention*], 61–71, Springer (2021). 1

[41] Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A. L., and Zhou, Y., "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306* (2021). 1

[42] Ji, Y., Zhang, R., Wang, H., Li, Z., Wu, L., Zhang, S., and Luo, P., "Multi-compound transformer for accurate biomedical image segmentation," in [*International Conference on Medical Image Computing and Computer-Assisted Intervention*], 326–336, Springer (2021). 1

[43] Chang, Y., Menghan, H., Guangtao, Z., and Xiao-Ping, Z., "Transclaw u-net: Claw u-net with transformers for medical image segmentation," *arXiv preprint arXiv:2107.05188* (2021). 1

[44] Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., and Wang, M., "Swin-unet: Unet-like pure transformer for medical image segmentation," *arXiv preprint arXiv:2105.05537* (2021). 1, 2

[45] Graham, B., El-Nouby, A., Touvron, H., Stock, P., Joulin, A., Jégou, H., and Douze, M., "Levit: a vision transformer in convnet's clothing for faster inference," in [*Proceedings of the IEEE/CVF international conference on computer vision*], 12259–12269 (2021). 1

[46] Xu, G., Wu, X., Zhang, X., and He, X., "Levit-unet: Make faster encoders with transformer for medical image segmentation," *arXiv preprint arXiv:2107.08623* (2021). 1

[47] Valanarasu, J. M. J., Oza, P., Hacihaliloglu, I., and Patel, V. M., "Medical transformer: Gated axial-attention for medical image segmentation," in [*International Conference on Medical Image Computing and Computer-Assisted Intervention*], 36–46, Springer (2021). 1

[48] Zhang, Y., Higashita, R., Fu, H., Xu, Y., Zhang, Y., Liu, H., Zhang, J., and Liu, J., "A multi-branch hybrid transformer network for corneal endothelial cell segmentation," in [*International Conference on Medical Image Computing and Computer-Assisted Intervention*], 99–108, Springer (2021). 1

[49] Wang, W., Chen, C., Ding, M., Yu, H., Zha, S., and Li, J., "Transbts: Multimodal brain tumor segmentation using transformer," in [*International Conference on Medical Image Computing and Computer-Assisted Intervention*], 109–119, Springer (2021). 1

[50] Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H. R., and Xu, D., "Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images," in [*International MICCAI Brainlesion Workshop*], 272–284, Springer (2022). 1

[51] Wu, Y., Liao, K., Chen, J., Chen, D. Z., Wang, J., Gao, H., and Wu, J., "D-former: A u-shaped dilated transformer for 3d medical image segmentation," *arXiv preprint arXiv:2201.00462* (2022). 1, 7

[52] Yu, W., Luo, M., Zhou, P., Si, C., Zhou, Y., Wang, X., Feng, J., and Yan, S., "Metaformer is actually what you need for vision," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*], (2022). 2, 3

[53] Tolstikhin, I. O., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J., et al., "Mlp-mixer: An all-mlp architecture for vision," *Advances in Neural Information Processing Systems* **34**, 24261–24272 (2021). 2

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

[55] Trockman, A. and Kolter, J. Z., "Patches are all you need?," *arXiv preprint arXiv:2201.09792* (2022). 2

[56] Armato III, S. G., Huisman, H., Drukker, K., Hadjiiski, L., Kirby, J. S., Petrick, N., Redmond, G., Giger, M. L., Cha, K., Mamonov, A., et al., "Prostatex challenges for computerized classification of prostate lesions from multiparametric magnetic resonance images," *Journal of Medical Imaging* **5**(4), 044501–044501 (2018). 3, 6

[57] Litjens, G., Toth, R., van de Ven, W., Hoeks, C., Kerkstra, S., van Ginneken, B., Vincent, G., Guillard, G., Birbeck, N., Zhang, J., et al., "Evaluation of prostate segmentation algorithms for mri: the promise12 challenge," *Medical image analysis* **18**(2), 359–373 (2014). 3, 6