3D Ultrasound Segmentation using Transformers

Ben Hers, Siyi Du
Department of Electrical and Computer Engineering, University of British Columbia
Vancouver, BC, Canada
{siyi, bhers}@ece.ubc.ca

Abstract

For many segmentation tasks the current state-of-the-art methods use standard encoder/decoder based architectures such as UNET[7]. One of the downsides of these convolutional based models is that they have limited receptive fields, so the model cannot look at all the relevant features in the image. Transformers split the images into patches and perform a projection on the image. Each transformed image patch is then treated as part of a sequence in natural language processing (NLP). This enables transformers to model long range dependencies or find global features by doing inter and intra patch learning in the images and achieve state-of-the-art performance. Transformers for vision applications normally require millions of training images to work well, but in scenarios like medical imaging, labeled data can be scarce. In this project we take 3D volumes from ultrasound scans and predict each voxel to be part of different bones. Since we have a limited dataset, we combine both aspects from CNNs and Transformers in a hybrid model which utilizes axial attention [13] to efficiently and effectively segment bones from ultrasound scans. By combining both techniques we were able to outperform all other state-of-the-art methods by 1.3%.

1 Introduction

As the computational capacity in graphics cards has been increasing, it has started to unlock the possibilities of training deep learning models on 3D volumes. Development dysplasia of the hip (DDH) is a disorder that can be screened using a 3D ultrasound scanner, and if caught early is treatable as an infant but if left unchecked can cause a variety of disorders later in life for the patients. Approximately 3% of all newborns have DDH [11] and it is estimated that $1.25 billion US per year [11] goes into the treatment of the consequences of DDH in adults, such as osteoarthritis. Researchers have been using convolutional neural network based models as the state-of-the-art for hip volume segmentation. Since data is hard to come by, the inductive bias such as translation equivariance and locality makes them easy to train and generalize [2] well to most problems. This has been especially useful in medical imaging scenarios where labeled data is scarce, so the inductive bias helps the network learn representations with limited data [2]. Current state-of-the-art techniques in ultrasound segmentation are based on the UNet++ architecture which is an encoder/decoder architecture with dense skip pathways [20]. One limitation of these convolutional neural network(CNN) architectures is that they cannot model long range interactions in the image, and in recent research we have seen transformer based methods outperform traditional CNNs in ultrasound tasks [13]. They show that important features may be spread across the entire image. CNN based methods have a fixed receptive field but transformers through inter patch attention can learn global features and outperform CNNs. This forces the CNN models to focus on local textures and patterns instead of considering the whole volume in its predictions [13]. Hybrid models that combine small convolutional based encoder/decoder with multi-head self attention layers in between the encoder/decoder has shown to be a promising direction. Using a combination of both multi-head self attention layers (MHSA) as well as convolutional encoders/decoders we can outperform CNN's and purely transformer based

1st EECE571F (2022).

methods. There has been research investigating these hybrid models such as UNETR with a purely transformer encoder [5] and COTR using MHSA layers in the bottleneck [17].
In this project we created a hybrid model that uses a small convolutional encoder and decoder with a transformer in the bottleneck which utilizes axial attention which was inspired by the MedTransformer [13] to create a model that outperforms not only classical CNN models like UNET but also hybrid models such as UNETR [5]. To achieve this we created our **H**ybrid model that leveraged **A**xial attention and used **R**esidual blocks[6] in the encoder and decoder for our **T**ransformer model giving us our **HART** model. Since our dataset was limited in size the convolutional encoder/decoders helped us learn relevant features and not overfit to our volumes. Axial Attention helped us reduce the complexity of the self attention layers for our large volumes and residual blocks helped with gradient propagation through the network. To summarize this paper:

1. We created a new hybrid architecture for 3D Ultrasound Volume Segmentation
2. Leveraged residual blocks as well as axial attention
3. Successfully improves the performance over both traditional convolutional networks as well as recent hybrid models.

2 Related Work

3D Ultrasound Segmentation The Biomedical Signal and Image Computing Labratory at UBC has pioneered developing 3D Ultrasound segmentation tools for detecting development dysplasia of the hip (DDH). Before the rise of deep learning, researchers initially focused on hand engineered features but Harimi et al. [4] was the first to develop a 3D UNet for DDH detection in infants. They utilized deep convolutional neural networks to get rid of hand engineering and improve the performance of the detection of DDH. To further improve the method, Kannan et al. [8] used bayesian techniques to add an uncertainty based loss term which helped the network improve its performance. This helps the model figure out which voxels it was uncertain about and use that in the loss function to influence the training.

Vision Transformers In 2020 Dosovitskiy et al. [2] were the first to apply a transformer from natural language processing [14] to vision with patches that were larger than 2x2. They showed that transformers could outperform CNN based architectures as long as they had large amounts of data. Following the original vision transformer, research has exploded [10][18] into how best to apply and leverage transformers and global attention in computer vision tasks. Transformers are able to focus on long range interactions inside of sentences [14] so by bringing the global attention into computer vision we are able to model interactions between features over the entire image. Traditional CNNs have a fixed receptive field and are forced to focus on local context and texture which does not let it incorporate features far away from each other in its decisions. Transformers do have their limitations, they lack the translation equivariance and locality [2] of CNNs which make CNNs easier to train. This makes the transformers much harder to train and require more data to achieve equivalent or better performance.

Transformers in Medical Imaging Since the original vision transformer required large amounts of data to achieve better performance, there has been research done into how to use transformers in medical imaging scenarios where data is scarce. One of the most promising directions has been the creation of hybrid transformer models that use convolutional encoders which then pass the feature maps into the transformer instead of the volume itself [17][3][1][19]. They then can do multi-head self attention on the feature maps which then get decoded through a series of convolutional layers. These models are able to leverage the translation equivariance and locality [2] of CNN's as well as the global attention of the transformer to help us outperform CNN models with small amounts of data. Hatamizadeh et al. [5] created a hybrid model that had a transformer based encoder with a convolutional based decoder for 3D brain tumour segmentation. Xie et al. was able to leverage deformable attention [17] inside the bottleneck of a UNET to outperform 3D UNET (CNN) based models. One limitation of transformers is that self attention is quadratic in complexity, Valanarasu et al. [13] leveraged axial attention [15] to reduce the complexity of attention in their model from quadratic to linear.

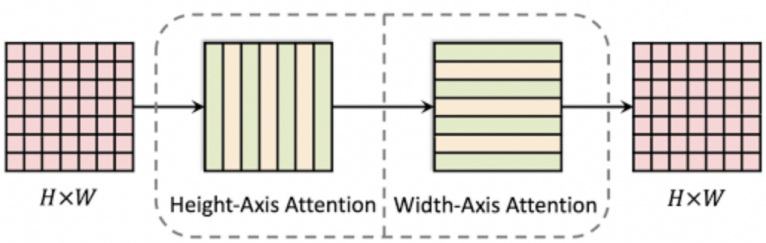

Figure 1: Axial Attention

3 Methods

Self Attention The basis for transformers is the multi-head self attention module which let us attend to different features in the feature maps by calculating the self attention using queries, keys, and values. We will start with a two dimensional feature map which we can represent as a tensor $x \in \mathbb{R}^{C_{in} \times H \times W}$ where C_{in} represents the number of feature maps and H, W is the height and width of the feature map. The resulting attention scores from the self attention can be calculated using the following equation:

$$z_{ij} = \sum_{h}^{H} \sum_{w}^{W} softmax(q_{ij} k_{hw}) v_{hw} \tag{1}$$

z is the output from the self attention block, q, k, and v are the queries, keys, and values of the attention block. To get the values across the feature maps, i spans from 1 to H and j spans from 1 to W. The values of the self attention are calculated using the global affinities from the softmax which is a matrix multiplication between the query and the keys [13]. This is a very powerful method for global attention that let us make predictions by comparing and focusing on features throughout the image versus just in a fixed receptive field. One downside of this is these matrix multiplications can be very costly and as the feature map dimensions or number of channels increase it can mean that we can no longer use self attention for images.

Axial Attention Calculating self-attention for volumes can be extremely computationally costly so to overcome this we can decompose the self attention to two self attention blocks that perform self attention on the height axis and width axis. We perform the attention on each axis independently and then combine them as shown in fig 1. This is known as axial attention and it models the original self attention more efficiently [13]. By decomposing the attention into two steps, the self attention becomes linear instead of quadratic with respect to the size of height and width of the input.

$$y_{ij} = \sum_{w=1}^{W} softmax(q_{ij}^T k_{iw} + q_{ij}^T r_{iw}^q + k_{iw}^T r_{iw}^k)(v_{iw} + r_{iw}^v) \tag{2}$$

This equation follows the equation in [15]. We can see in equation 2 the equation for axial attention in the width dimension, for our applications we perform the same steps for height and depth in the volume. This method of doing the height axis then the width axis efficiently [13] models the original self attention.

Model overview We have introduced our **H**ybrid**A**xial**R**esidual**T**ransformer which utilizes residual blocks [6] with skip connections as shown in Fig 2. Each blue block in Fig 2 represents a residual block that uses 3D convolutional layers with internal skip connections. Our **HART** model utilizes these residual blocks to increase the dimensionality in the channel dimension but reduces our volume size to $\frac{H}{8}$ by $\frac{W}{8}$ by $\frac{D}{8}$ where H is the Height of the volume, W is the width, and D is the depth. We then pass a tensor $x \in \mathbb{R}^{B \times C \times \frac{H}{8} \times \frac{W}{8} \times \frac{D}{8}}$ where C is the channel dimension in the final residual layer of the encoder and B is the batch size. Just as in natural language processing [14], transformers get a 1D sequence as input and we perform self attention on these sequences as described above. To get from our 3D volumes to 1D sequences, we pass in our 3D volume as 3D patches or cubes which we then flatten and pass through a linear transformation.

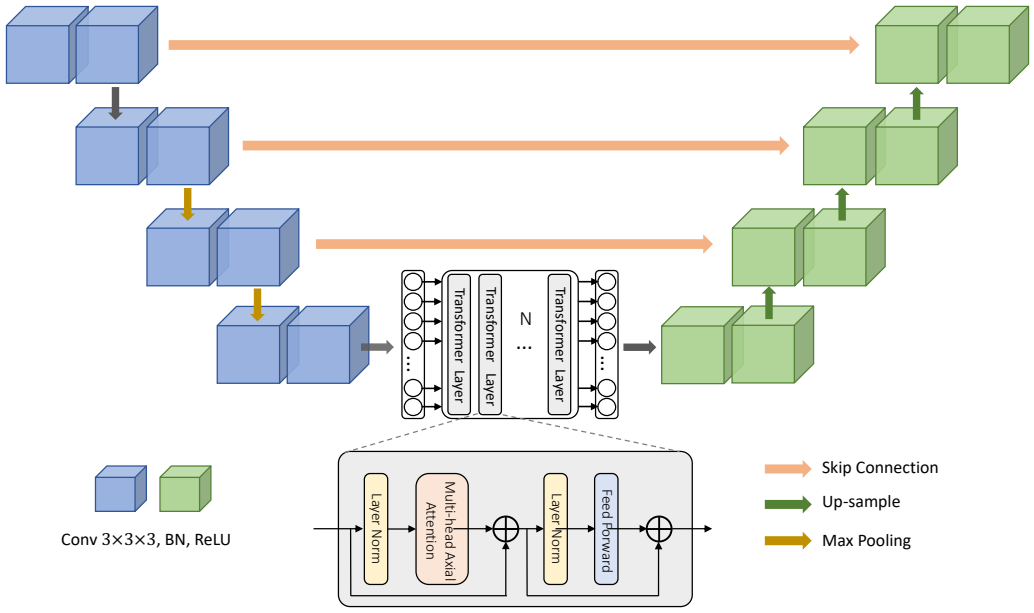

Figure 2: Model Architecture

Hybrid Model Following papers such as TransBTS [16], we created a hybrid model that had a convolutional encoder and decoder with transformer in the middle. Our model differs where we incorporate both axial attention and residual blocks into the architecture whereas most hybrid models such as TransBTS [16] used VGG like blocks. To further improve the model performance, we utilize skip connections between the encoder and decoder so the model could look at features from both the encoder and decoder together which let the decoder learn how to incorporate global information from the transformer and local information from the encoder.

4 Experiments

To compare our model we use data from our lab at the Biomedical Signal and Image Computing Laboratory at UBC. We have a partnership with Children's Hosiptal in Vancouver where we take 3D ultrasound scans of infants age 0-6 months where we screen patients for DDH. Labelled data is very scarce since doctors need to label 3D volumes, so we are limited to 63 individual patients where we have 180 scans coming from those patients. To split up the data we split the 63 patients into 80% training(50) and 20% testing(13) to avoid training and testing on the same patient. This gives us around 142 3D volumes for testing and 38 for testing. We trained all our models on a single TITAN V until the validation loss and metrics converge. One of the main limiting factor is that we are limited to a batch size of 1 since these GPU's have a memory of 12GB. The nature of our data is that they are decently large volumes, so we could only fit a single sample on our GPU at once. This meant we used instance normalization [12]. With small batch sizes BatchNorm can be unstable or unlearnable since we cannot get an accurate representation of the parameters of our data.

Experimental Setup To help this we resized all our volumes to a size of $128 \times 128 \times 128$ to let us fit our models into the gpu memory. Due to the time frame of the project we use binary cross entropy loss since our labels have a single class and trained with the Adam optimizer [9]. We used a learning rate of 0.001 and trained all the models for 50 epochs or 49.90k steps.

5 Results

5.1 Evaluation Metrics

To evaluate our models we use the F1 score and Dice score and Intersection over Union(IoU) which are good metrics to measure our segmentation accuracy. The F1 score is the harmonized mean of both precision and recall which gives a good single score to compare models since some models have higher recall and lower precision than other models and vice versa. Precision can be calculated by all the true positives (where both prediction and label are positive) divided by true positives and false positives (where prediction is 1 but label is 0). This tells us out of all our positive predictions, how many are correct. Recall can be calculated by all the true positives divided by true positives and false negatives (where prediction is 0 but label is 1). Precision only tells us what ratio of our positives are correct, whereas recall tells us out of all the positives in the label how many did we accurately predict. F1 score is a combination of these 2 and can be calculated as below.

$$F1 = 2 \times \frac{p \times r}{p + r} \tag{3}$$

where p is the precision score between 0 and 1 and r is the recall score between 0 and 1. As we can see this F1 metric gives us a good way to look at the combination of precision and recall in a single metric. Dice score is extremely similar but is calculated slightly differently.

$$Dice = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{4}$$

where TP is true positives, FP is false positives, and FN is false negatives. As we can see both of these are extremely similar so they give us a good score for our segmentation accuracy.
The other score we measure is the IoU or Jaccard Index which measures how much our predicted segmentation map overlaps with the true segmentation map. It measures the number of voxels that are in common between the model and label, which can be calculated as below,

$$IoU = \frac{model \cup label}{model \cap label} \tag{5}$$

5.2 Quantitative Results

We run multiple trials for every model and measure all the metrics mentioned above. To compare our hybrid models against state-of-the-art convolutional models we compare our model against 3D UNet and 3D UNet++ [20]. These methods use standard encoder/decoder based architectures to learn a higher dimensional representation of the image and then transform the representation into heatmaps. UNet++ [20][7] added in dense skip pathways to help increase the performance of the model. UNETR[5] is a hybrid model created by MONAI and Nvidia that has a purely transformer based encoder but using different convolutional layers in the decoder to perform segmentation. The ResHybrid model is simply our **HART** model without the axial attention so we can see the results of using axial attention in our model. The combination of residual encoder and decoder blocks as well as global attention allowed our model to outperform the other models. The axial attention [13] allowed our model to have 2 million more parameters than the largest CNN based method, which gave our model more capacity to learn. Combining a better architecture and more capacity to learn allowed us to get the best results.

Metric	**HART**	3D UNET	3D UNET++	UNETR	ResHybrid
F1	**87.03±0.33%**	85.76±0.2%	83.30±0.61%	69.97±0.07%	83.86±0.03%
Dice	**85.3±0.06%**	83.95±0.16%	82.20±0.54%	68.10±0.25%	82.20±0.11%
IOU	**76.38±0.29%**	74.07±0.23%	72.09±0.70%	54.29±0.29%	73.79±1.61%
Recall	**87.62±0.24%**	84.54±0.28%	84.79±0.65%	68.84±1.22%	83.70±0.64%
Precision	86.83±1.01%	**87.76±0.75%**	82.88±0.64%	73.57±2.24%	86.03±0.54%

Table 1: Performance of CNN and Hybrid Models on 3D Ultrasound Data

5.3 Qualitative Results

To compare our **HART** model against the other models we load the best weights for each of the models and look at the predicted heatmaps out. Since we cannot show animations through the volume we show representative slices of the volume. We can see in Figure 3 we compare the two best models:

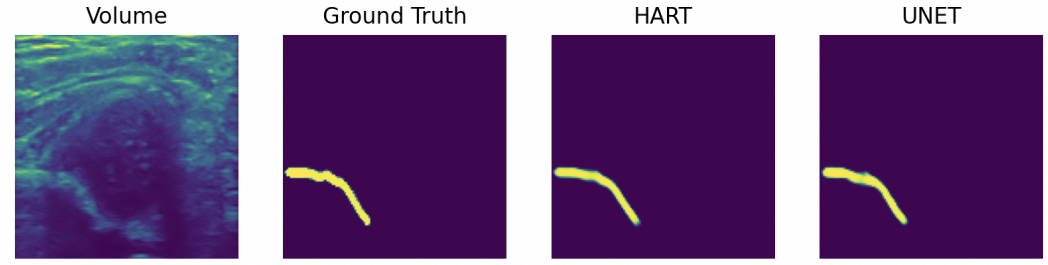

Figure 3: Visualizing model performance

Our **HART** model and 3D UNET model. We can see that our model has more consistent predictions, whereas the 3D UNET model has spikes in along the boundaries of the bone. The global attention allows the model to look at the entire bone structure to generate a smooth curve in the bone.

5.4 Model size

To have a better picture at the reasons for the differing performances of the model, the number of training parameters or model parameters can be a good indicator. We can see that the two best performing models have the most parameters so this could be an indicator that since we use 3D volumes our models cannot fully learn the representations as limited by our GPU. One reason we can see in Table 2 that the ResHybridViT model does not perform well is because we could not fit enough parameters in the GPU due to the model architecture.

Model	Training Parameters
UNET	7.9 million
UNET++	7.0 million
HART	10 million
ResHybridViT	2.6 million

Table 2: Comparing number of training parameters

6 Ablation Studies

Since the novel contribution of this project is using transformers inside of UNET based architectures, we run ablation studies to test the model without transformers and then also with and without axial attention. The first row of Table 3 is our model but we take out all of the transformer or MHSA layers.

Model	Dice score
No Transformer or Axial Attention	81.83%
Transformer but no axial attention	82.20%
Transformer and Axial Attention	85.3%

Table 3: Ablation Study of effect of attention layers

This shows us that the addition of MHSA layers into a UNET-like structure shows a modest improvement. This is because we can perform global attention on the feature maps that allow the model to look further than the local receptive field of each pixel. An interesting development is that due to

the efficiency of axial attention we are able to greatly increase performance. We believe this may be due to the implementation of the axial attention because if we look at table 2 we can see our **HART** model has 10 million parameters but the normal attention variant has just 2.6 million parameters. That increased number of parameters could be the largest difference in performance. Future work should perform further ablation studies where we artificially limit the number of parameters in the **HART** or we use clusters like Compute Canada with larger GPUs to increase the number of parameters in the normal attention model. Due to the time limitation of the project we were unable to do this.

7 Discussion

We show that for our lab's 3D ultrasound data, using a hybrid model has the best test performance because purely transformer based models as mentioned in the original ViT[2] paper needed large amounts of data and pre-training. By incorporating encoders and decoders that leverage residual blocks [6] and skip connections our model is able to use the invariance and ease of training of convolutional models while also using the global attention in the multi-head attention blocks in the bottleneck of our hybrid model.
One interesting thing we note is that the simple 3D UNET outperforms both the 3D UNet++ as well as UNETR. One challenging aspect of our data is the lack of data so UNETR struggles to learn the best representation of our data. Since we were time limited in this project, we were unable to run extremely long training runs on our data to try to let UNETR perform better, so with more time and fine-tuning it would be interesting to compare UNETR performance.
When comparing our models, we noticed that introducing axial attention into our network greatly increased performance over the normal ResHybrid transformer model that had a standard transformer based bottleneck. Since we were limited in GPU size, we found that we had to decrease the number of parameters in the network to get even a single volume to fit in our GPU and have enough memory to store the gradients to update the network. This highlights the benefit of axial attention as it is more efficient and leads to performance differences due to the number of parameters in the TITAN V and architecture of the network.

8 Future Work

One area for improvement was the loss function of the model. We simply used a binary cross entropy loss for our model since we outputted for every voxel a probability between 0 and 1. The volume is mostly background and a small portion is the bone so investigating loss functions like focal losses or even dice losses could be an interesting direction to improve performance. Also due to the time limitations of the project we wanted to benchmark our model against more state of the art models like TransBTS since as our model evolved it was quite similar to TransBTS [16].
Our convolutional encoders and decoders go from the original volume dimensions to $1/8$ of the original size. One potential area for improvement would be to add in additional layers to reduce the volume size even more and add more skip connections between the encoder and decoder. Currently there is two concatenations that go between the encoder and decoder but creating deeper encoders would allow more skip connections. This could be done to tune the best network architecture for our task but could be complicated by our limited amount of data. A solution for this would be to find other open source medical imaging datasets and creating training, validation, and testing sets and perform a neural architecture search.

9 Conclusion

In this paper we present a new architecture for 3D Ultrasound Segmentation that builds off recent research in both medical imaging and transformers. Our **HART** model is a hybrid model that has a convolutional encoder and decoder that leverages both residual blocks in the convolutional blocks as well as multi-head attention layers in the bottleneck of the model. Combining these techniques, we were able to outperform prior research done in 3D Ultrasound Segmentation for the detection of developmental dyplasia of the hip. We also outperformed UNETR [5] which is a state-of-the-art hybrid architecture. Overall, our hybrid model can help improve the performance of hip bone segmentation and help doctors screen infants for DDH.

10 Author Contribution

Siyi Implemented the UNET models and some of the ViT original models we started out with. Wrote related works section, and experiments section. Also helped with writing UNET specific parts of the methods/discussion.

Ben Implemented the residual blocks and built up the infrasture using pytorch lightning. Also wrote the axial attention and hybrid model. Wrote the introduction, methods, and discussion.

Combined We both contributed to the code base, and writing report was joint effort with a lot of back and forth on writing discussion, future work, and abstract.

References

[1] 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". In: *arXiv preprint arXiv:2102.04306* (2021).

[2] 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". In: *arXiv preprint arXiv:2010.11929* (2020).

[3] 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*. Springer. 2021, pp. 61–71.

[4] Houssam El-Hariri, Antony J Hodgson, Kishore Mulpuri, and Rafeef Garbi. "Automatically Delineating Key Anatomy in 3-D Ultrasound Volumes for Hip Dysplasia Screening". In: *Ultrasound in Medicine and Biology* 47.9 (2021), pp. 2713–2722.

[5] 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, pp. 574–584.

[6] 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*. 2016, pp. 770–778.

[7] Huimin Huang, Lanfen Lin, Ruofeng Tong, Hongjie Hu, Qiaowei Zhang, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, and Jian Wu. "Unet 3+: A full-scale connected unet for medical image segmentation". In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2020, pp. 1055–1059.

[8] Arunkumar Kannan, Antony Hodgson, Kishore Mulpuri, and Rafeef Garbi. "Leveraging voxel-wise segmentation uncertainty to improve reliability in assessment of paediatric dysplasia of the hip". In: *International Journal of Computer Assisted Radiology and Surgery* 16.7 (2021), pp. 1121–1129.

[9] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[10] 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 (ICCV)*. Oct. 2021, pp. 10012–10022.

[11] Randall T Loder and Elaine N Skopelja. "The epidemiology and demographics of hip dysplasia". In: *International Scholarly Research Notices* 2011 (2011).

[12] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. "Instance normalization: The missing ingredient for fast stylization". In: *arXiv preprint arXiv:1607.08022* (2016).

[13] Jeya Maria Jose Valanarasu, Poojan Oza, Ilker Hacihaliloglu, and Vishal M Patel. "Medical transformer: Gated axial-attention for medical image segmentation". In: (2021), pp. 36–46.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[15] Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. "Axial-deeplab: Stand-alone axial-attention for panoptic segmentation". In: *European Conference on Computer Vision*. Springer. 2020, pp. 108–126.

[16] 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*. Springer. 2021, pp. 109–119.

[17] Yutong Xie, Jianpeng Zhang, Chunhua Shen, and Yong Xia. "Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation". In: (2021), pp. 171–180.

[18] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis E.H. Tay, Jiashi Feng, and Shuicheng Yan. "Tokens-to-Token ViT: Training Vision Transformers From Scratch on ImageNet". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2021, pp. 558–567.

[19] Yundong Zhang, Huiye Liu, and Qiang Hu. "Transfuse: Fusing transformers and cnns for medical image segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2021, pp. 14–24.

[20] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. "Unet++: A nested u-net architecture for medical image segmentation". In: (2018), pp. 3–11.