Enhanced Kidney Tumor Segmentation in CT Scans Using a Simplified UNETR with Organ Information

Sanghyuk Roy Choi, Kanghyeok Ko, Sun Jae Baek, Soyeon Lee
Department of Intelligent Semiconductor Engineering
Chung-Ang University
Seoul, Korea
{choiroy, dogworld12, baechu, soyeon1608}@cau.ac.kr

Jungro Lee, Minhyeok Lee*
School of Electrical and Electronics Engineering
Chung-Ang University
Seoul, Korea
{jungro7982, mlee}@cau.ac.kr

Abstract—**The rising incidence of cancer diagnoses necessitates efficient tumor detection methods in CT scans. Manual tumor identification by physicians is labor-intensive and demands high level of focus. To address these challenges, we introduce a deep learning model for automated tumor detection. Our model employs a streamlined version of the U-Net Transformer (UNETR), where the original transformer layers are replaced by Squeeze and Excitation (SE) layers for more efficient computation. This modification improves the Dice score for tumor segmentation and enhances the ability to distinguish between organ and tumor pixels. Furthermore, we establish that concurrent segmentation of both the organ and the tumor significantly improves the overall performance in tumor segmentation tasks. To evaluate this claim, we trained the model using two types of datasets: one containing both organ and tumor information, and another containing only tumor information. The former approach yielded more accurate tumor localization, while the latter proved ineffective due to the absence of organ context. Our findings suggest that incorporating organ information significantly improves the training and prediction accuracy for tumor segmentation.**

Keywords—organ segmentation, tumor segmentation, medical segmentation, deep learning, Squeeze and Excitation network, Transformer

I. INTRODUCTION

Numerous Convolution Neural Network (CNN) based models conducting medical image segmentation tasks have been proposed [1-4]. For example, U-Net segments organs in Computed Tomography (CT) scan using encoder and decoder architecture [5]. The encoder extracts features from CT scan, while the decoder reconstructs the output image leveraging the feature representations received from the encoder. Also, UNET++ is proposed for the medical segmentation tasks which has nested architecture that every encoder and decoder layers are connected through dense connection [6]. With the dense connections, UNET++ can integrate small features from the input image. Moreover, CNN based architecture denoted as KiU-NET has been proposed [7]. The KiU-NET has two different branches in parallel which are U-Net and Kite-Net. Motivated from U-Net, Kite-Net upsamples feature to high dimensional feature space and downsamples the feature, thereby facilitating the e acquisition of relatively small features. Consequently, KiU-Net conducts detailed prediction with two networks.

Inspired by development in Natural Language Processing (NLP), the transformer is applied to vision models. Recently, the appearance of Vision Transformers (ViT) that adopts the transformer has promoted significant advancement in computer vision field surpassing the CNN-based models [8-12]. The feature extraction through the transformer improves generalization ability to various datasets and has made significant advance in image segmentation by fusing the patch information. Motivated from ViT, UNet Transformer (UNETR) utilizes an advantage of the transformer. Multi-Head Self Attention (MHSA) layer in transformer gradually extracts the information among patches which is transmitted to the decoder subsequently. Lastly, the decoder inherits the patch information and reconstructs image through deconvolution operations.

In this paper, we propose a lightweight UNETR conducting kidney and kidney tumor segmentation task from the CT scans. We enhance the efficiency of the UNETR architecture by replacing the MHSA layer with the Squeeze and Excitation (SE) layer, which computes the attention among feature channels [13-18]. As a result, The SE layer facilitates the capacity to rank the feature channels.

We conduct experiments using Kidney Tumor Segmentation Dataset (KiTS19) which includes the CT scans with the annotation of kidney and kidney tumor. We train our model using a dataset that has both kidney and kidney tumor information. Then, we compare accuracy between our proposed model and a baseline model being trained with a different dataset that contains only tumor information.

II. RELATED WORKS

A. Squeeze and Excitation Network

CNNs have demonstrated their efficiency in addressing vision tasks dominating the deep learning field that conducts vision tasks [1-4]. CNN layer identifies localized spatial features across input channels and treats input channels equally, without considering attention between input channels. However, SE network examines attention of input channels to perceive interdependency between feature channels.

SE network has two operations that are squeeze operation and excitation operation. Squeeze operation is executed by global average pooling. This operation entails the compression of input features, effectively combining each 2D feature channel into a single vector. The excitation operation generates a vector for input channels. Once the vector is determined, the vector is channel-wisely multiplied with input channels scaling the importance between input channels, thereby establishing a rank among the input channels.

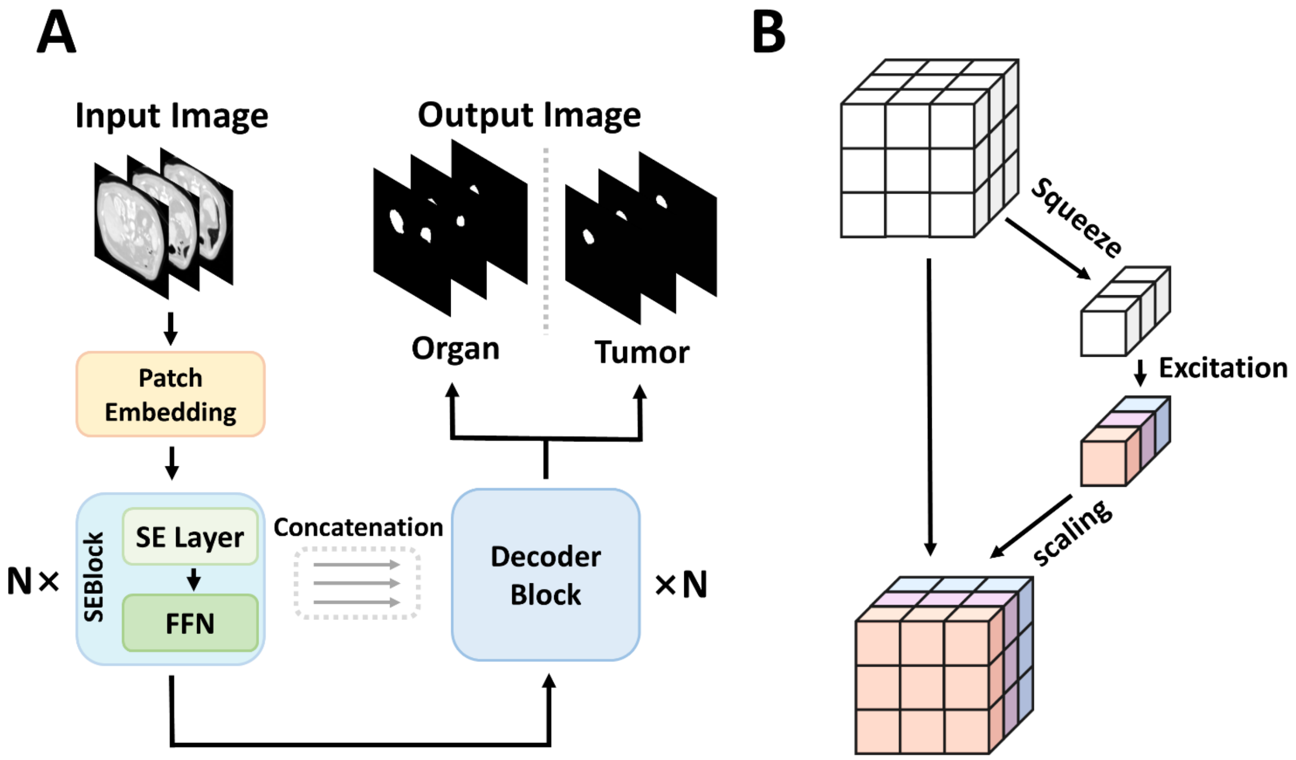

Fig 1. The structure of our proposed model. (**A**) represents overall architecture of the model. (**B**) is the structure of SE layer. One column vector implies attention between 2D channels.

B. U-Net

The CNN-based U-Net is initially designed for biomedical image segmentation [5]. The U-Net architecture consists of contracting path and expansive path. The contracting path captures the context and compress the input image information. The expansive path which is symmetric with respect to the contracting path enables precise feature localization. The expansive path upsamples feature maps until the final output dimension of U-Net is identical to input dimension.

C. Vision Transformer

Transformers are originally introduced within the domain of machine translation. Subsequently, transformers emerged as dominant models for NLP tasks [19, 20]. Inspired by the success of transformer in NLP tasks, the researchers attempt to integrate transformer into CNNs to enhance the performance in computer vision tasks [8, 9, 21, 22]. The ViT is one of the first attempt that employs pure transformer-based architecture, demonstrating competitive performance against CNNs in the domain of image classification tasks. Unlike CNNs, ViT based models possess the capacity to capture the global contextual information by means of patch images comparisons.

D. UNet Transformer

UNETR is designed to perform 3D medical segmentation tasks [23-25]. Motivated from ViT, UNETR leverages Multi-Head Self Attention (MHSA) layer to inherit the advantages. MHSA layer performs feature analysis among 3D patches that are uniformly split from the original input image.

UNETR is comprised of two parts which are encoder, decoder. The encoder extracts patch information across multiple channels and produces feature channels, employing a sequence of MHSA and MLP layers. The decoder is comprised of 3D deconvolution and convolution layers that contribute to the output image reconstruction. The skip connection between the encoder and decoder is employed to convey extracted information from the encoder to the decoder at various layers. With these components, UNETR proficiently generates the desired output.

E. Kidney and Kidey Tumor Segmentation Dataset

KiTS19 is a CT scan dataset that has a CT image and annotation of right and left kidneys and kidney tumors. The annotation process has been done by medical students under supervision of a professional urologic oncologist. The dimensions of slices are identical for every slice. However, the volumes differ for each CT scan that minimum number of slices in the CT scan is 29 and the maximum is 1059. KiTS19 dataset is available at https://github.com/neheller/kits19.

III. Methods

In the absence of organ context, accurate prediction of tumor localization becomes notably challenging. The dimension of tumor is considerably small compared to the

CT scan and kidney, making the tumor detection demanding. Consequently, providing the organ and tumor locations concurrently to deep learning model helps the model to locate tumors in detail. Additionally, we propose modified UNETR model that MHSA layers in UNETR are replaced with SE layers.

An overview of the proposed model is described in Figure 1. Our proposed model treats 3D CT scans as an input. Treating every pixel in the CT scans causes substantial computational complexity. Therefore, inspired by UNETR, we utilize patch embedding layer that compress input image into a vector. The patch embedding layer divides an input image ($x \in R^{H \times W \times D}$) into 3D patches ($x_p \in R^{N \times P^3}$) and $N = HWD/P^3$ is number of patches. The patch embedding layer transforms patch sequences into tokens. Due to the lack of positional information in tokens, a learnable positional encoding vector is added to tokens [26, 27]. The embedding equation can be represented as

$$z_{tokens} = [p_1E; p_2E; \cdots; p_NE] + E_{pos} \qquad (1)$$

where z_{tokens} is token vector which is transformed from patches; p_n is a patch vector; E is a matrix that transforms patch vectors into tokens; E_{pos} is a learnable positional encoding vector that injects positional information.

UNETR utilizes MHSA layers to evaluate the relevance between the patches. However, the computational complexity of MHSA layer grows exponentially as the resolution of image increases. To resolve the increasing computational complexity, we employ SE layer instead of MHSA layer. The SE layer focuses on important features which is similar role of MHSA layer. Adopting SE layer, we build hierarchical Squeeze and Excitation Block (SEBlock) that SE layers and MLP layers are connected sequentially. We construct an encoder architecture with 2, 4 and 6 SEBlocks connected successively. The outputs from [2nd, 4th, 6th] SEBlocks are concatenated to decoder layers to fuse extracted features between the encoder and decoder.

To maintain weight values stable in SEBlock, the layer normalization is applied for every layer. The operation of the SEBlock can be represented by the following equations:

$$S_l = LN(MLP(LN(v_{l-1} \times S_{l-1}))), \qquad (2)$$

where S_n is input of SE layer; LN is layer normalization; MLP is MLP layer; v_n is a vector that injects attention into feature map channels.

The decoder block consists of 3D deconvolutional layers to reconstruct images from the feature maps. The decoder block increases the resolution of feature maps through deconvolutional operation. This process is repeated until the resolution of image is identical to input resolution. At the output stage, the two output channels predict the location of organ and tumor individually.

For the loss calculation and model performance evaluation, the Dice coefficient is used [28-30]. The formula of Dice coefficient can be represented by the following equations:

$$S_{dice} = \frac{2\times(P_{true}\times P_{pred})}{P_{true}+P_{pred}}, \qquad (3)$$

where S_{dice} indicates Dice coefficient; P_{true} is a binary matrix that has location information of organ or tumor [28]. In the datasets, the value one indicates organ and tumor pixels and zero indicates the background; P_{pred} is a binary matrix predicted by the model, indicating the locations of the organ and tumor; The total Dice loss formula is represented as follows:

$$S_{dice,total} = 0.65 \times S_{dice,organ} + 0.35 \times S_{dice,tumor}, \qquad (4)$$

where $S_{dice,total}$ is total Dice coefficient that dice coefficients of organ and tumor are added in the ratio of 65 and 35; $S_{dice,organ}$ is Dice coefficient calculated between the predicted organ segmentation and ground truth; $S_{dice,tumor}$ is Dice coefficient calculated between the predicted tumor segmentation and ground truth.

The two Dice losses are calculated for organ and tumor prediction independently and added with different weights. For Dice loss of tumor prediction, the weight value of 0.35 is applied since the size of tumors is considerably small compared to CT scan and kidney. The total Dice loss is minimized by backpropagation algorithm [33].

IV. Results

We train our model with KiTS19 dataset for medical segmentation task. KiTS19 has 544 3D CT scans which are annotated by medical students under the supervision of the author of a paper describing KiTS19. The CT scans consist of 29 to 1059 slices that have 512 × 512 pixels for a single slice. To overcome the lack of computational resources, we reshaped the dataset into 128 × 128 × 128 using linear interpolation during pre-processing. Decreasing the size of CT scans causes tumor pixels to combine with neighboring pixels resulting tumor pixels to vanish. Therefore, we excluded 54 datasets that do not contain tumor pixels. The min-max scaling is applied to normalize the values in each dataset.

For the kidney segmentation task, 489 CT volumes with kidney body and kidney tumor annotations are used. We split the dataset into training set, validation set and test set at ratio 70:20:10. The model is trained with AdamW optimizer applying the uniform learning rate of 0.0001 [31]. In addition, we employ data augmentation technique that is rotating images in range 0 to 10 degrees [32].

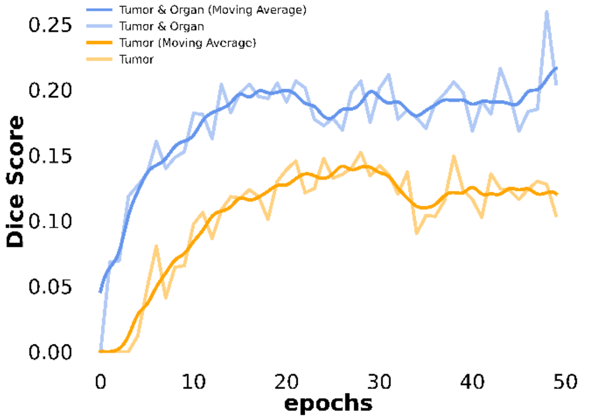

Fig 2. The Dice scores of our proposed model. The blue colored lines are Dice scores that trained with both organ and tumor dataset. The orange colored lines are Dice scores that trained with only tumor dataset. For the dark colored lines, moving average is applied.

The performance of our proposed model is estimated with Dice score. There are two types of models. Our model has two channels which learns the location of organ and tumor. However, the baseline model has only one channel that learns only the location of tumor. We compare the Dice scores from two models. Our model outperforms for 0.0711 percent point compared to the baseline model. As a result, the support of the organ information contributes to the prediction of tumor.

MHSA uses key, query, and value matrices to evaluate attention between the feature channels resulting huge computational burden. The SE layer is similar to MHSA layer that evaluates the attention between the feature channels. By adapting SE layer in place of MHSA layer, the computational complexity decreases by 13.9 percent and still can estimate the importance among the feature channels.

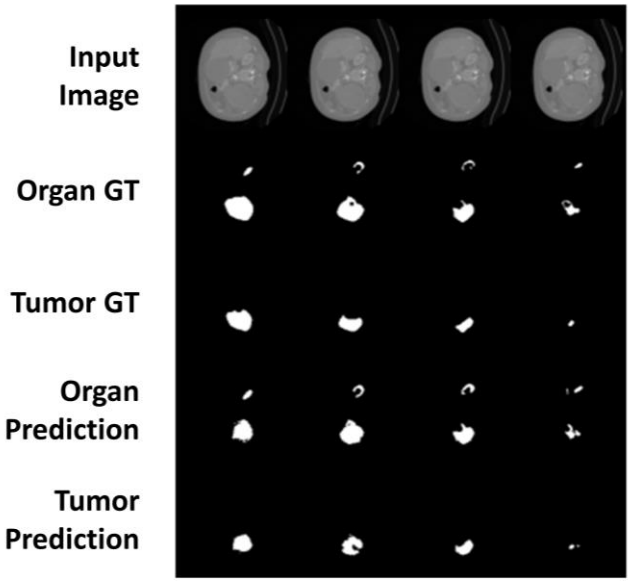

Fig 3. The organ and tumor prediction of our proposed model. The first row images are input images which are the CT scans. The second row images are the organ ground truth. The third row images are the tumor ground truth. The fourth row images are the organ predictions. The fifth row images are the tumor predictions.

V. Conclusion

This paper proposed a modified UNETR that MHSA layers are substituted with SE layers. MHSA utilized key, query and value matrices to calculate the attention of feature channels requiring substantial calculation. However, SE layer had only one vector for calculating the attention. As a result, the computational complexity decreased by 13.9 percent.

Our proposed model captured kidney and kidney tumor simultaneously. Also, our model performed improved segmentation capacity on tumor compared to segmenting tumor solely. Owing to providing the kidney information to tumor segmentation channel through backpropagation, the Dice score on tumor was enhanced by 0.0711 percent point. Our results demonstrated that the injection of organ information segments more precisely compared to segmenting kidney tumor without the organ information.

However, the study has its limitation. The dataset used for training the model comprises only 489 CT images annotated by medical students, which is not sufficiently diverse to generalize the model performance. Given the limitation, future research should aim to expand the dataset to include a more diverse range of CT images, annotated by experienced radiologists. The expansion of the dataset could provide a more robust evaluation of the model performance.

Moreover, future research could attempt combining additional features, such as patient history or other biomarkers, to consider diverse medical factors. Advanced techniques like generative adversarial network or diffusion model could also be employed to improve the robustness against variations in tumor segmentation.

Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00251528), and Korea Institute for Advancement of Technology (KIAT) grant funded by the Korea Government (MOTIE) (P0020967, Advanced Training Program for Smart Sensor Engineers)

References

[1] Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., and Cai, J.: 'Recent advances in convolutional neural networks', Pattern recognition, 2018, 77, pp. 354-377

[2] Albawi, S., Mohammed, T.A., and Al-Zawi, S.: 'Understanding of a convolutional neural network', in Editor (Ed.)^(Eds.): 'Book Understanding of a convolutional neural network' (Ieee, 2017, edn.), pp. 1-6

[3] Li, Q., Cai, W., Wang, X., Zhou, Y., Feng, D.D., and Chen, M.: 'Medical image classification with convolutional neural network', in Editor (Ed.)^(Eds.): 'Book Medical image classification with convolutional neural network' (IEEE, 2014, edn.), pp. 844-848

[4] O'Shea, K., and Nash, R.: 'An introduction to convolutional neural networks', arXiv preprint arXiv:1511.08458, 2015

[5] Ronneberger, O., Fischer, P., and Brox, T.: 'U-net: Convolutional networks for biomedical image segmentation', in Editor (Ed.)^(Eds.): 'Book U-net: Convolutional networks for biomedical image segmentation' (Springer, 2015, edn.), pp. 234-241

[6] Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., and Liang, J.: 'Unet++: A nested u-net architecture for medical image

segmentation', in Editor (Ed.)^(Eds.): 'Book Unet++: A nested u-net architecture for medical image segmentation' (Springer, 2018, edn.), pp. 3-11

[7] Valanarasu, J.M.J., Sindagi, V.A., Hacihaliloglu, I., and Patel, V.M.: 'Kiu-net: Overcomplete convolutional architectures for biomedical image and volumetric segmentation', IEEE Transactions on Medical Imaging, 2021, 41, (4), pp. 965-976

[8] Zhou, D., Kang, B., Jin, X., Yang, L., Lian, X., Jiang, Z., Hou, Q., and Feng, J.: 'Deepvit: Towards deeper vision transformer', arXiv preprint arXiv:2103.11886, 2021

[9] Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., and Shah, M.: 'Transformers in vision: A survey', ACM computing surveys (CSUR), 2022, 54, (10s), pp. 1-41

[10] Yao, T., Li, Y., Pan, Y., Wang, Y., Zhang, X.-P., and Mei, T.: 'Dual vision transformer', IEEE transactions on pattern analysis and machine intelligence, 2023

[11] Chen, J., He, Y., Frey, E.C., Li, Y., and Du, Y.: 'Vit-v-net: Vision transformer for unsupervised volumetric medical image registration', arXiv preprint arXiv:2104.06468, 2021

[12] Lee, S.H., Lee, S., and Song, B.C.: 'Vision transformer for small-size datasets', arXiv preprint arXiv:2112.13492, 2021

[13] Hu, J., Shen, L., and Sun, G.: 'Squeeze-and-excitation networks', in Editor (Ed.)^(Eds.): 'Book Squeeze-and-excitation networks' (2018, edn.), pp. 7132-7141

[14] Cheng, X., Li, X., Yang, J., and Tai, Y.: 'SESR: Single image super resolution with recursive squeeze and excitation networks', in Editor (Ed.)^(Eds.): 'Book SESR: Single image super resolution with recursive squeeze and excitation networks' (IEEE, 2018, edn.), pp. 147-152

[15] Roy, S.K., Dubey, S.R., Chatterjee, S., and Baran Chaudhuri, B.: 'FuSENet: fused squeeze-and-excitation network for spectral-spatial hyperspectral image classification', IET Image Processing, 2020, 14, (8), pp. 1653-1661

[16] Gu, J., Sun, X., Zhang, Y., Fu, K., and Wang, L.: 'Deep residual squeeze and excitation network for remote sensing image super-resolution', Remote Sensing, 2019, 11, (15), pp. 1817

[17] Choi, S.R., and Lee, M.: 'Estimating the prognosis of low-grade glioma with gene attention using multi-omics and multi-modal schemes', Biology, 2022, 11, (10), pp. 1462

[18] Lee, M.: 'An ensemble deep learning model with a gene attention mechanism for estimating the prognosis of low-grade glioma', Biology, 2022, 11, (4), pp. 586

[19] Kalyan, K.S., Rajasekharan, A., and Sangeetha, S.: 'Ammus: A survey of transformer-based pretrained models in natural language processing', arXiv preprint arXiv:2108.05542, 2021

[20] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., and Funtowicz, M.: 'Transformers: State-of-the-art natural language processing', in Editor (Ed.)^(Eds.): 'Book Transformers: State-of-the-art natural language processing' (2020, edn.), pp. 38-45

[21] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., and Gelly, S.: 'An image is worth 16x16 words: Transformers for image recognition at scale', arXiv preprint arXiv:2010.11929, 2020

[22] d'Ascoli, S., Touvron, H., Leavitt, M.L., Morcos, A.S., Biroli, G., and Sagun, L.: 'Convit: Improving vision transformers with soft convolutional inductive biases', in Editor (Ed.)^(Eds.): 'Book Convit: Improving vision transformers with soft convolutional inductive biases' (PMLR, 2021, edn.), pp. 2286-2296

[23] 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 Editor (Ed.)^(Eds.): 'Book Unetr: Transformers for 3d medical image segmentation' (2022, edn.), pp. 574-584

[24] Tao, H., Mao, K., and Zhao, Y.: 'DBT-UNETR: Double Branch Transformer with Cross Fusion for 3D Medical Image Segmentation', in Editor (Ed.)^(Eds.): 'Book DBT-UNETR: Double Branch Transformer with Cross Fusion for 3D Medical Image Segmentation' (IEEE, 2022, edn.), pp. 1213-1218

[25] Chu, H., De la O Arévalo, L.R., Tang, W., Ma, B., Li, Y., De Biase, A., Both, S., Langendijk, J.A., van Ooijen, P., and Sijtsema, N.M.: 'Swin UNETR for Tumor and Lymph Node Segmentation Using 3D PET/CT Imaging: A Transfer Learning Approach': '3D Head and Neck Tumor Segmentation in PET/CT Challenge' (Springer, 2022), pp. 114-120

[26] Chen, P.-C., Tsai, H., Bhojanapalli, S., Chung, H.W., Chang, Y.-W., and Ferng, C.-S.: 'A simple and effective positional encoding for transformers', arXiv preprint arXiv:2104.08698, 2021

[27] Wang, Y.-A., and Chen, Y.-N.: 'What do position embeddings learn? an empirical study of pre-trained language model positional encoding', arXiv preprint arXiv:2010.04903, 2020

[28] Ghosal, S., Xie, A., and Shah, P.: 'Uncertainty quantified deep learning for predicting dice coefficient of digital histopathology image segmentation', arXiv preprint arXiv:2109.00115, 2021

[29] Bertels, J., Eelbode, T., Berman, M., Vandermeulen, D., Maes, F., Bisschops, R., and Blaschko, M.B.: 'Optimizing the dice score and jaccard index for medical image segmentation: Theory and practice', in Editor (Ed.)^(Eds.): 'Book Optimizing the dice score and jaccard index for medical image segmentation: Theory and practice' (Springer, 2019, edn.), pp. 92-100

[30] Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., and Jorge Cardoso, M.: 'Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations', in Editor (Ed.)^(Eds.): 'Book Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations' (Springer, 2017, edn.), pp. 240-248

[31] Loshchilov, I., and Hutter, F.: 'Decoupled weight decay regularization', arXiv preprint arXiv:1711.05101, 2017

[32] Shijie, J., Ping, W., Peiyi, J., and Siping, H.: 'Research on data augmentation for image classification based on convolution neural networks', in Editor (Ed.)^(Eds.): 'Book Research on data augmentation for image classification based on convolution neural networks' (IEEE, 2017, edn.), pp. 4165-4170

[33] Rumelhart, D.E., Hinton, G.E., and Williams, R.J.: 'Learning representations by back-propagating errors', nature, 1986, 323, (6088), pp. 533-536