2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI) | 978-1-6654-2923-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISBI52829.2022.9761417

SEGTRANSVAE: HYBRID CNN - TRANSFORMER WITH REGULARIZATION FOR MEDICAL IMAGE SEGMENTATION

Quan-Dung Pham[⋆] *Hai Nguyen-Truong*[⋆†‡] *Nam Nguyen Phuong*[⋆]
Khoa N. A. Nguyen[⋆†‡] *Chanh D. T. Nguyen*[⋆§], *Trung Bui*, *Steven Q.H. Truong*[⋆]

[⋆] VinBrain JSC., Vietnam [†] University of Science, Ho Chi Minh City, Vietnam
[‡] Vietnam National University, Ho Chi Minh City, Vietnam § Vin University, Vietnam

ABSTRACT

Current research on deep learning for medical image segmentation exposes their limitations in learning either global semantic information or local contextual information. To tackle these issues, a novel network named SegTransVAE is proposed in this paper. SegTransVAE is built upon encoder-decoder architecture, exploiting transformer with the variational autoencoder (VAE) branch to the network to reconstruct the input images jointly with segmentation. To the best of our knowledge, this is the first method combining the success of CNN, transformer, and VAE. Evaluation on various recently introduced datasets shows that SegTransVAE outperforms previous methods in Dice Score and 95%-Haudorff Distance while having comparable inference time to a simple CNN-based architecture network. The source code is available at: `https://github.com/itruonghai/SegTransVAE`.

***Index Terms*—** Transformer, Variational Autoencoder, Medical Image Segmentation, MRI brain tumor, CT kidney.

1. INTRODUCTION

Since the introduction of U-Net [1], many state-of-the-art deep neural networks for medical image segmentation have been proposed. CNN-based segmentation networks such as U-Net [1], and SegresnetVAE [2] are developed on a symmetric encoder-decoder architecture with skip connection, which combines high resolution features from the contracting path with the upsampled output. Then, this information can then be learned by a successive convolution layer to assemble a more precise output. However, they pose their limitation on learning global context and long-range spatial dependencies. As a result, this raises challenges to learn global semantic information which plays a critical role in segmentation tasks.

Transformer-based models in the natural language processing (NLP) domain have achieved state-of-the-art results. Inspired by attention mechanisms [3] in NLP, recent research such as UNETR [4] surpasses the aforementioned limitation in segmentation task by exploiting this mechanism. The self-attention mechanism in the transformers enables them to dynamically highlight the crucial features of sequences and learn their long-range dependencies. UNETR [4] leverages the power of transformers for volumetric medical image segmentation. A pure transformer is utilized as the encoder to learn contextual information from the embedded input patches. The extracted representations from the transformer encoder are merged with a decoder via skip connections at multiple resolutions to predict segmentation outputs. However, local structures are ignored when directly splitting images into patches as tokens for transformer, as mentioned in the research of Yuan et al. [5]. Moreover, UNETR [4] lacks inductive bias such as translation equivariance and locality, and therefore does not generalize well when trained on insufficient amounts of data.

In this work, a novel network named SegTransVAE is proposed to complement the drawbacks of existing studies. SegTransVAE is built upon an encoder-decoder architecture with the variational autoencoder (VAE) branch as the encoder regularization to the network to reconstruct the input images jointly with segmentation. Thanks to VAE branch, the proposed network can avoid the overfitting problem. First, the encoder of the network uses 3D CNN to extract the volumetric spatial features and downsample the input 3D images, which effectively captures the local 3D context information. Second, each volume is reshaped into a vector and fed into the transformer encoder for global feature modeling. Third, the 3D CNN decoder takes the feature embedding from transformer and performs progressive upsampling while the extracted representations from the encoder are concatenated with a decoder via skip connections at multiple resolutions to predict segmentation outputs.

2. METHOD

The architecture of the proposed method is shown in Fig.1. This approach follows encoder-decoder architecture with an asymmetrically larger encoder to extract image features, the

Quan-Dung Pham and Hai Nguyen-Truong equally contributed.

transformer encoder to model the long-distance dependency in a global space and a smaller decoder to construct the segmentation mask. Also, an additional VAE branch is added to the endpoint of the transformer to reconstruct source images.

2.1. Encoder component

Inspired by ResNet [6], in this research, a modified-Resnet block is proposed in which consists of two convolutions with instance normalization [7] and Leaky ReLU, followed by additive identity skip connection. This modified-ResNet block suffers from sparse gradients and shows a significant qualitative improvement. The encoder part uses the proposed modified-ResNet blocks. In order be able to model the image local context information across spatial and depth dimensions for volumetric segmentation, the modified-ResNet blocks are stacked with downsampling to gradually encode input $X \in \mathbb{R}^{C \times H \times W \times D}$ images into low-resolution and high-level feature representation $F \in \mathbb{R}^{K \times \frac{H}{8} \times \frac{W}{8} \times \frac{D}{8}}$. After that, this representation is fed into the transformer encoder to further learn long-range correlations with a global receptive field.

2.2. Transformer component

2.2.1. Feature embedding

A linear projection is used to project the feature map F from K dimensions to a d dimensional embedding space f in order to ensure a comprehensive representation of each volume. In order to encode the location information, the learnable position embeddings [8] are used and fused with the $d \times N$ feature map f by direct addition, where $N = \frac{H}{8} \times \frac{W}{8} \times \frac{D}{8}$. This creates the feature embeddings as follows:

$$z_0 = f + PE = W \times F + PE, \tag{1}$$

where the linear projection matrix is W, the position embeddings is $PE \in \mathbb{R}^{d \times N}$, and the feature embeddings is $z_0 \in \mathbb{R}^{d \times N}$.

2.2.2. Transformer encoder

A stack of transformer layers [9] is utilized to construct transformer encoder in which each transformer layer consisting of Multi-Head Attention (MHA) and Feed Forward Network (FFN) sublayers according to

$$z'_l = \text{MHA}(\text{LN}(z_{l-1})) + z_{l-1}, \tag{2}$$

$$z_l = \text{FFN}(\text{LN}(z'_l)) + z'_l, \tag{3}$$

where LN refers to the layer normalization and z_l denotes the output of l-th transformer layer.

2.2.3. Feature mapping

A feature mapping module is added to project the sequence data back to a standard feature map. Then, this feature map is fed as the input dimension of 3D CNN decoder. In feature mapping module, the output sequence of transformer is $z_L \in \mathbb{R}^{d \times N}$ is first reshaped into $d \times \frac{H}{8} \times \frac{W}{8} \times \frac{D}{8}$ then a convolution block is employed to reduce the channel dimension from d to K. Finally a feature map $Z \in \mathbb{R}^{K \times \frac{H}{8} \times \frac{W}{8} \times \frac{D}{8}}$ is obtained.

2.3. Decoder component

The encoder also uses modified-ResNet blocks to perform feature upsampling and pixel-level segmentation, but with a single block per spatial level. Each decoder level begins with an upsizing to reduce the number of features by a factor of 2 and double the spatial dimension, followed by a concatenation of encoder output of the equivalent spatial level. The end of the decoder has the same spatial size as the original image and the number of features equal to the initial input feature size, followed by $1 \times 1 \times 1$ convolution into 3 channels and a sigmoid function.

2.4. VAE component

Variational autoencoder (VAE) is added to reconstruct the volumetric input segmentation. The main role of VAE branch is to avoid the overfitting problem and to increase the network generalization. From the encoder endpoint output, the input is reduced to a lower-dimensional space of 256 in which 128 represents for mean, and the rest represents for standard deviation. A sample is drawn from the Gaussian distribution with the given mean and standard deviation $\mathcal{N}(\mu, \sigma^2)$, then reconstructed into the input image dimensions following the same architecture as the decoder.

2.5. Loss Function

Let y and $\hat{y}$ be the ground truth of segmentation and the prediction of the model, respectively. To avoid training data having no label as $\hat{y} = y = 0$, ε is added into numerator and denominator. Dice Loss is as follows

$$\mathcal{L}_{\text{Dice}}(y, \hat{y}) = 1 - \frac{2\hat{y}y + \varepsilon}{\hat{y} + y + \varepsilon}. \tag{4}$$

VAE loss is a total loss of reconstruction loss on VAE $\mathcal{L}_{\text{Rec}}$ branch and standard VAE penalty term $\mathcal{L}_{\text{KL}}$. Let $x_{\text{reconstruction}}$ and x denote the reconstruction image and input image, respectively.

In this study, $\mathcal{L}_{\text{Rec}}$ is the mean square error over each voxels:

$$\mathcal{L}_{\text{Rec}} = \|x_{\text{reconstruction}} - x\|_2^2. \tag{5}$$

$\mathcal{L}_{\text{KL}}$ is a Kullback–Leibler divergence between the estimated normal distribution $\mathcal{N}(\mu, \sigma^2)$ and a prior distribution

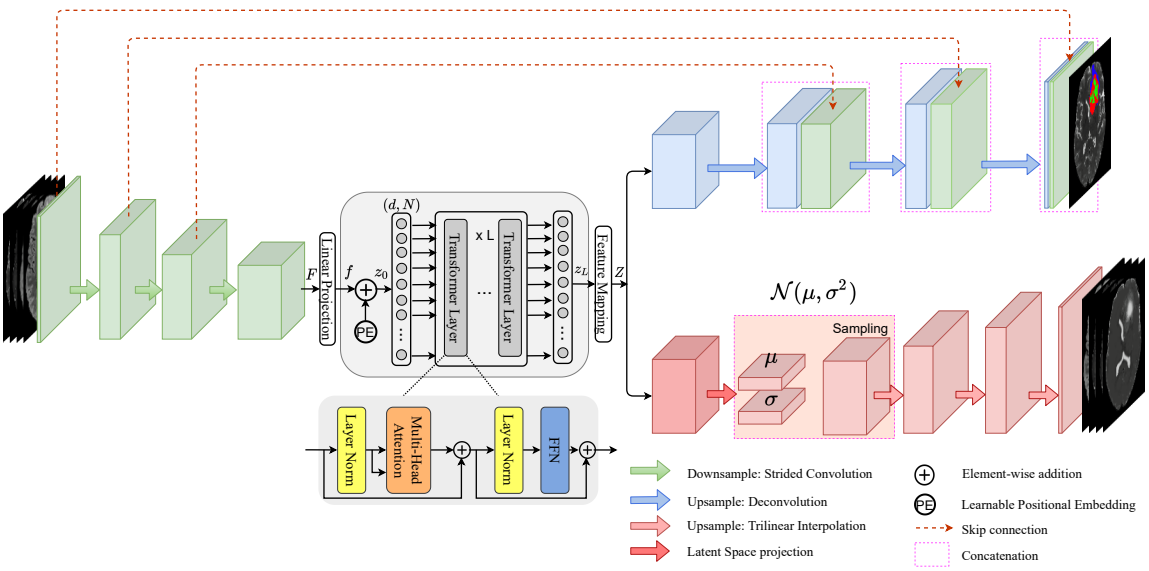

Fig. 1: Overview architecture of proposed method.

$\mathcal{N}(0, 1)$ as

$$\mathcal{L}_{\text{KL}} = \frac{1}{N_{\text{total voxels}}} \sum \mu^2 + \sigma^2 - \log \sigma^2 - 1, \qquad (6)$$

where $N_{\text{total voxels}}$ is the total number of image voxels.

The final loss function is the combination of Dice Loss and VAE Loss as follow

$$\mathcal{L} = \mathcal{L}_{\text{Dice}} + 0.1 \times (\mathcal{L}_{\text{Rec}} + \mathcal{L}_{\text{KL}}). \qquad (7)$$

A hyper-parameter (regularization factor weight) of 0.1 is chosen to provide a good balance between dice loss and VAE loss as [2].

3. EXPERIMENT

3.1. Experimental Setup

3.1.1. Dataset

The proposed method is evaluated on newly introduced BraTS 2021 [10] and KiTS19 [11]. BraTS 2021 [10] provides a 3D brain MRI dataset with tumor segmentation labels annotated. The training dataset comprises 1251 cases for training and 219 for validation rigidly aligned and resampled to a uniform isotropic resolution of $1mm^3$. The input image size is $240 \times 240 \times 155$. The KiTS19 [11] dataset is a collection of segmented CT imaging and treatment outcomes for 300 patients treated with partial or radical nephrectomy between 2010 and 2018.

Since the validation data of BraTS 2021 is private and it is not provided the ground truth, in this evaluation, 1251 cases is split as 1000 cases for training/validation and 251 cases for testing. Due to the small number of training images in KiTS19 [11], five-fold cross-validation is chosen to evaluate proposed method and conventional models on this dataset. During training, the BraTS 2021 [10] input images are cropped of size $128 \times 128 \times 128$ while KiTS19 [11] input images are cropped of size $128 \times 160 \times 256$.

3.1.2. Evaluation Metrics

The metrics Dice score and 95% - Hausdorff distance (HD) are used for quantitative evaluation.

3.2. Quantitative Results

3.2.1. BraTS 2021

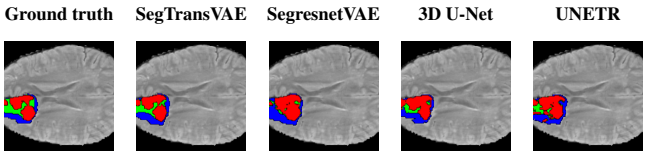

Fig. 2: The visual comparison of BraTS segmentation results where red, green, blue are the enhancing tumor, core tumor and whole tumor, respectively.

In this experiment, the proposed method SegTransVAE is compared with state-of-the-art 3D approaches including 3D U-Net [12], UNETR [4], and SegresnetVAE [2]. Table 1 illustrates the Dice Score comparison between SegTransVAE and previous methods. It is clear that SegTransVAE outperforms previous research as it achieves the Dice Score of 85.48%, 90.42%, and 92.60% on ET, WT and TC, respectively. In

Table 1: Dice Score and 95%-HD comparison on BraTS.

Method	Dice Score (%)			95%-HD (mm)		
	ET	**WT**	**TC**	**ET**	**WT**	**TC**
3D U-Net [12]	80.69	85.00	90.11	4.83	6.20	8.99
UNETR [4]	82.18	85.14	89.46	5.63	7.62	13.18
SegresnetVAE [2]	84.46	89.52	92.35	3.24	3.79	6.36
SegTransVAE	**85.48**	**90.52**	**92.60**	**2.89**	**3.57**	**5.84**

terms of 95% - Hausdorff Distance, Table 1 shows that SegTransVAE also achieves considerable improvement. It is clear that due to leveraging CNN for high-level features extracting and transformer for global feature modeling, the proposed method shows its significant improvement in segmentation. It is obvious that in Fig. 2 SegTransVAE creates segmentation masks of brain tumors more precisely and especially generates much better segmentation masks of the small tumor as enhancing tumor.

3.2.2. KiTS 2019

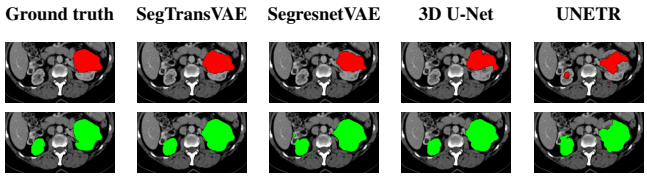

Fig. 3: The visual comparison of KiTS segmentation results where red and green are tumor and kidney, respectively.

Table 2: Dice Score and 95%-HD comparison of Kidney.

Method	**Kidney**	
	Dice Score (%)	95%-HD (mm)
3D U-Net [12]	92.37 ± 4.54	6.32 ± 1.83
UNETR [4]	91.86 ± 1.29	8.84 ± 1.78
SegresnetVAE [2]	94.86 ± 0.69	4.28 ± 0.97
SegTransVAE	$\mathbf{95.28 \pm 0.85}$	$\mathbf{3.28 \pm 1.19}$

Table 3: Dice Score and 95%-HD comparison of Tumor.

Method	**Tumor**	
	Dice Score (%)	95%-HD (mm)
3D U-Net [12]	60.41 ± 4.14	42.02 ± 4.05
UNETR [4]	34.87 ± 3.80	60.43 ± 9.43
SegresnetVAE [2]	63.67 ± 4.39	25.86 ± 3.24
SegTransVAE	$\mathbf{66.31 \pm 4.41}$	$\mathbf{24.61 \pm 2.49}$

The proposed method is also evaluated on KiTS 2019 dataset [11]. Tables 2 and 3 illustrate that SegTransVAE outperforms in tumor segmentation and shows comparable results in kidney segmentation of the conventional methods as 3D U-Net [12], UNETR [4], and SegresnetVAE [2]. In addition, the proposed method shows better results in kidney and tumor in every fold of the experiment. By utilizing VAE, SegTransVAE shows its significant results in the little availability of training data as KiTS 2019 dataset [11]. It is clear that in Fig. 3, SegTransVAE shows better performance in segmentation tumor and kidney.

3.3. Complexity

Table 4: Comparision of number of parameters and averaged inference time.

Method	#Params (M)	Inference Time (s)
3D U-Net [12]	5.6	0.45
UNETR [4]	101.7	0.38
SegresnetVAE [2]	7.5	0.55
SegTransVAE	44.7	0.45

The complexity of SegTransVAE is compared to other models in terms of the number of parameters and the averaged inference time. The benchmark is calculated based on the input size of $(4, 128, 128, 128)$. Table 4 illustrates that SegTransVAE has 44.7M parameters as compared to 101.7M parameters of UNETR [4] which makes [4] hard to converge, especially with high-resolution input. As a result, the proposed method outperforms [4] at all evaluation metrics on BraTS 2021 and KiTS19 datasets. Although CNN-based segmentation methods as 3D U-Net and SegresnetVAE [2] have fewer parameters than UNETR [4] and SegTransVAE, the GFLOPs benchmarks of CNN-based methods are more than UNETR and SegTransVAE, with the GFLOPs benchmarks of 3D U-Net and SegresnetVAE are more than 1000 GFLOPs while those of UNETR and SegTransVAE are 358.8 GFLOPs and 607.5 GFLOPs, respectively. As a consequence, SegTransVAE is less complex than the CNN-based network. Moreover, SegTransVAE has the second-lowest averaged inference time after UNETR and is comparable to simple CNN-based architecture like 3D U-Net. Also, SegTransVAE is 20% faster than SegresnetVAE.

4. CONCLUSION

A novel network named SegTransVAE is presented with the goal of complement the disadvantages of existing studies and the little availability of training data. SegTransVAE is built upon encoder-decoder architecture with the variational autoencoder (VAE) branch to the network to reconstruct the input images jointly with segmentation. transformer is also used for global feature modeling. Experiments on two distinct datasets demonstrate the superiority of the proposed method when compared to state-of-the-art methods including 3D U-Net [12], UNETR [4], and SegresnetVAE [2].

5. REFERENCES

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

[2] Andriy Myronenko, "3d mri brain tumor segmentation using autoencoder regularization," in *International MICCAI Brainlesion Workshop*. Springer, 2018, pp. 311–320.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[4] Ali Hatamizadeh, Dong Yang, Holger Roth, and Daguang Xu, "Unetr: Transformers for 3d medical image segmentation," *arXiv preprint arXiv:2103.10504*, 2021.

[5] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zihang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan, "Tokens-to-token vit: Training vision transformers from scratch on imagenet," *arXiv preprint arXiv:2101.11986*, 2021.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Identity mappings in deep residual networks," in *European conference on computer vision*. Springer, 2016, pp. 630–645.

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

[8] 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.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[10] Ujjwal Baid, Satyam Ghodasara, Michel Bilello, Suyash Mohan, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C. Kitamura, Sarthak Pati, Luciano M. Prevedello, Jeffrey D. Rudie, Chiharu Sako, Russell T. Shinohara, Timothy Bergquist, Rong Chai, James Eddy, Julia Elliott, Walter Reade, Thomas Schaffter, Thomas Yu, Jiaxin Zheng, BraTS Annotators, Christos Davatzikos, John Mongan, Christopher Hess, Soonmee Cha, Javier E. Villanueva-Meyer, John B. Freymann, Justin S. Kirby, Benedikt Wiestler, Priscila Crivellaro, Rivka R. Colen, Aikaterini Kotrotsou, Daniel S. Marcus, Mikhail Milchenko, Arash Nazeri, Hassan M. Fathallah-Shaykh, Roland Wiest, András Jakab, Marc-André Weber, Abhishek Mahajan, Bjoern H. Menze, Adam E. Flanders, and Spyridon Bakas, "The RSNA-ASNR-MICCAI brats 2021 benchmark on brain tumor segmentation and radiogenomic classification," *CoRR*, vol. abs/2107.02314, 2021.

[11] Nicholas Heller, Niranjan Sathianathen, Arveen Kalapara, Edward Walczak, Keenan Moore, Heather Kaluzniak, Joel Rosenberg, Paul Blake, Zachary Rengel, Makinna Oestreich, et al., "The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes," *arXiv preprint arXiv:1904.00445*, 2019.

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