SegMAE-Net: A Hybrid Method Using Masked Autoencoders for Consistent 3D Medical Image Segmentation

Zheng Kai Liaw[1], Ankit Das[2], Shaista Hussain[2], Feng Yang[2], Yong Liu[2], Rick Siow Mong Goh[2]
[1]*Department of Mathematics,National University of Singapore*, Singapore
[2]*Institute of High Performance Computing*
Agency for Science, Technology and Research
Singapore 138632, Republic of Singapore
[1]liawzhengkai@u.nus.edu
[2]{dasak, hussains, yangf, liuyong, gohsm}@ihpc.a-star.edu.sg

***Abstract*—Volumetric medical image segmentation has been particularly challenging as both local and global features are important in producing an accurate and consistent segmentation output. However, 2D CNNs often ignore the global contextual information of the volumetric input while the use of 3D CNNs is heavily limited by the large computational needs and GPU restrictions. In this paper, we propose SegMAE-Net, which combines 2D and 3D methods to leverage on the strengths of both approaches. Specifically, SegMAE-Net consists of two branches, 1) slice-centric branch made up of an encoder-decoder architecture to learn the local features of each image slice, 2) volume-centric branch utilising masked autoencoders to capture long-range dependencies. We evaluate SegMAE-Net on the RETOUCH dataset, and the experimental results show that our proposed method achieves state-of-the-art performance. We also show that our method is able to produce segmentation outputs with a higher consistency across the volume level.**

***Index Terms*—medical image segmentation, masked autoencoders, convolutional neural networks**

I. Introduction

Medical image segmentation is an important procedure in many medical applications, such as computer-aided diagnosis and clinical interventions [1]–[3]. This is because a region of interest is extracted in these images to provide a comprehensive analysis of certain anatomical structures through automatic or manual means. However, the segmentation of 3D volumetric data has been particularly challenging, due to the variation in imaging, anatomy and pathology. Two main approaches are usually adopted in the segmentation of such volumetric data. One such method involves splitting the volumetric data into a stack of 2D slices and segmenting each slice individually. This process can be tedious, considering the large amount of data to segment. While automatic methods can alleviate the manual process, such solutions do not consider the global context within the image volume, potentially leading to inconsistencies in the segmentation result between adjacent slices. The other approach uses the volumetric input directly to produce the segmentation output. However, such an approach is often limited by the GPU storage and memory. To counter this, the volume may be downsampled before performing convolution, which may result in loss of local features.

Convolutional neural networks (CNNs) have proven to be more accurate and effective in various medical image segmentation and health case-based tasks [4], [5], as compared to traditional image segmentation methods, including edge detection-based [6] and region-based methods [7]. Particularly, U-Net, originally developed for 2D biomedical image segmentation, received much attention due to its encoder-decoder architecture and skip connections [8]. Since then, several variations of U-Net have emerged to address the limitations of the original U-Net. While volumetric data cannot be directly processed using the conventional U-Net, each slice can be segmented individually and then stacked together to form a 3D segmentation output. However, such an approach fails to consider the inter-slice information and thus, the segmentation output may appear inconsistent in its overall shape. 3D U-Net solves this problem by performing 3D convolution operations directly on a volumetric input [9]. V-Net also proposes a similar approach as 3D U-Net except for its use of residual connections [10]. Several attempts have also been made to combine the slice-centric approach (2D only) with the volume-centric approach (3D only) for segmentation of medical images. SA-Net proposes a hybrid segmentation approach with the use of 3D convolutions and a dimension reduction mechanism to convert 3D information learnt in the reconstruction branch into 2D information to aid segmentation [11]. SCAA also adopts a similar method that uses self-attention mechanism instead to generate an aggregated 2D feature map from the 3D feature maps [12]. Transformers [13] and vision transformers [14] have also shown much potential in the field of medical image segmentation due to their effectiveness in modelling long-range dependencies. Particularly, UNETR [15] and TransUNet [16] both utilise self-attention mechanisms to capture global

This research received support from the Agency for Science, Technology and Research (A*STAR) AME Programmatic Funds (Grant Number: A20H4b0141).

979-8-3503-5409-6/24/$31.00 ©2024 IEEE
DOI 10.1109/CAI59869.2024.00225

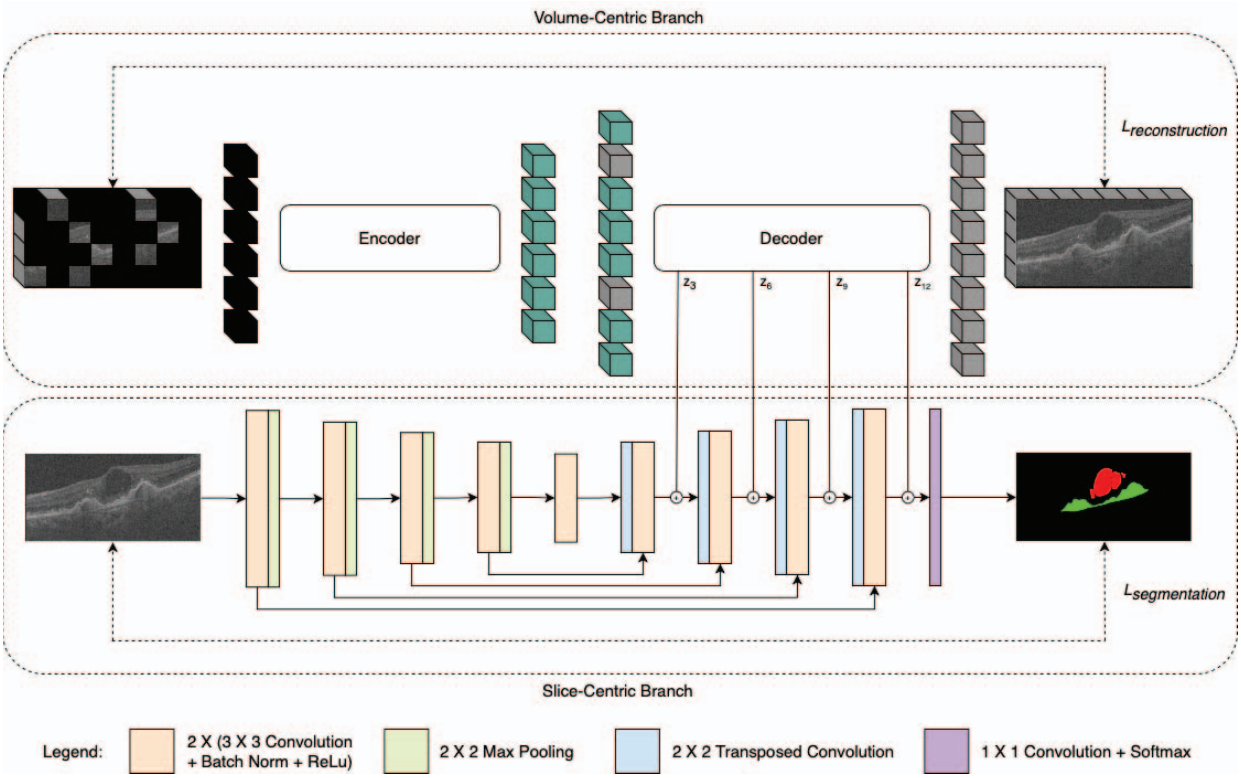

Fig. 1. Overview of our proposed SegMAE-Net, which consists of a slice-centric branch and a volume-centric branch. The feature aggregation branch is not shown in this figure, please refer to Fig. 2 for more details.

information of the embedded 3D image patches in volumetric medical image segmentation. [17] similarly proposes a causal knowledge fusion framework consisting of 3D hierarchial attention mechanisms to tackle the challenge of cross-modality volumetric segmentation. Recently, masked autoencoders have also gained popularity in image reconstruction, due to their improved capabilities in learning useful information with less domain knowledge and improving generalisability [18].

In this work, we propose to incorporate volume-centric information into the slice-centric segmentation of 3D medical images through the use of masked autoencoders. Specifically, we introduce **SegMAE-Net**, a hybrid 2D and 3D approach that learns the volume-centric information between adjacent slices and aggregates the knowledge to inform and improve the segmentation of a particular slice. We hypothesise that the entire volume input is often not needed in capturing the global volumetric context, especially for large input volumes. Instead, we prove that a certain number of adjacent slices in the forward and backward directions would be sufficient in achieving the same purpose. In summary, the contributions of this paper are:

1) We propose a novel architecture that utilises both masked autoencoders and convolutions to perform volumetric segmentation. The use of masked autoencoders to guide the segmentation of 3D medical images further establishes its effectiveness in different downstream tasks.
2) Our proposed SegMAE-Net achieves a notable improvement in the segmentation accuracy on the RETOUCH dataset as compared to other state-of-the-art models.

II. Method

A. Overview

As discussed earlier, both slice-centric and volume-centric information are important to obtain a precise and consistent segmentation result. To this end, we propose a hybrid network that aims to integrate both slice-centric and volume-centric information to produce the segmentation output. The network consists of two branches: a slice-centric branch that utilises an encoder-decoder architecture with repeated convolution and up-sampling to learn the local features of each slice, and a volume-centric branch that utilises masked autoencoders to reconstruct a masked volume made up of several adjacent slices to learn the contextual information of the adjacent slices. These two branches are connected by a feature aggregation branch to consolidate the information learnt in the volume-centric branch into the slice-centric branch to produce the segmented slice output. Each segmented slice is then stacked together depth-wise to form the volumetric segmentation output. The overall architecture is shown on Fig. 1.

B. Volume-Centric Branch

Our volume-centric branch is built upon masked autoencoders (MAE) to provide useful information about the 3D contextual information of adjacent slices to the slice-centric branch. Specifically, given a particular slice S_i of the volumetric data, the volume-centric branch will take in a volume inclusive of its n adjacent slices in both the forward and backward directions ($\{S_{i-n}, \ldots, S_i, \ldots, S_{i+n}\}$) to perform the reconstruction, where n is the number of adjacent slices that is used. For slices with insufficient neighbouring slices, we replace the missing slices with the mean pixel value of the available slices. Only the visible patches are mapped to a latent space through the MAE encoder, and the pixel values of the masked patches would then be predicted with the MAE decoder. The key elements of MAE are described below:

1) Input: The input volume $x \in \mathbb{R}^{H \times W \times C}$ is first reshaped to a sequence of $N = HW/P^2$ non-overlapping patches $x_p \in \mathbb{R}^{N \times (P^2 C)}$, where N is the number of patches, (H, W) is the resolution of each image slice, C is the total number of slices used in the input volume (i.e., $C = 2n + 1$) and (P, P) is the patch size. The patches are then linearly embedded, and positional embedding of these patches is added to capture the positional information.

2) Masking: A subset of the patches is masked and removed, while the remainder is fed into the MAE encoder. The random sampling technique is used as our masking strategy, with a sufficiently high masking ratio to ensure that the task would not be solved simply by extrapolation of unmasked neighbouring patches.

3) MAE Encoder: The MAE encoder is a vanilla ViT architecture applied only on unmasked patches. This significantly reduces the computational time and memory. After being fed into the MAE encoder, the unmasked patches are then mapped into the latent space.

4) MAE Decoder: The MAE decoder takes in the encoded features of unmasked patches in the latent space and the mask token as inputs, where the mask token is a learnable and shared vector. A series of transformer blocks are adopted as the decoder.

5) Reconstruction: Finally, the MAE predicts the pixel values for each masked patch to reconstruct the input volume. The mean squared error between the original and reconstructed image is then computed as the loss function, though this is only applied to the masked patches. Specifically, let $y^{input} \in \mathbb{R}^{CHW \times 1}$ represent the input pixel values and $y^{pred} \in \mathbb{R}^{CHW \times 1}$ represent the predicted pixel values. The reconstruction loss can then be written as

$$L_{reconstruction} = \frac{1}{\Omega(y_M^{input})} \sum_{i \in M} (y_i^{pred} - y_i^{input})^2 \quad (1)$$

where M represents the set of masked pixels, i represents the pixel index and $\Omega(\cdot)$ represents the cardinality of the set.

C. Slice-Centric Branch

Our slice-centric branch follows the U-Net architecture [8]. This branch only takes in one image slice S_i to evaluate the segmentation output. The encoder path consists of multiple blocks of two 3×3 convolutional layers and one 2×2 max-pooling layer to learn the spatial features of the volume slice and reduce the resolution map by half respectively. Batch normalisation is also implemented to speed up and stabilise the training process [19]. In the decoder path, each up-sampling block consists of a 2×2 transposed convolutional layer and two 3×3 convolutional layers. To prevent the loss of information from layer to layer, skip connections are also implemented between the contracting and expansive path. The output from each feature aggregation block is also summed with the output of each up-sampling block using element-wise addition to integrate the local features of each slice with its contextual knowledge. The final layer comprises of a 1×1 convolutional layer and a softmax activation layer. We used the dice loss and cross-entropy loss as the segmentation loss between each 2D segmented output and the ground truth slice, specifically

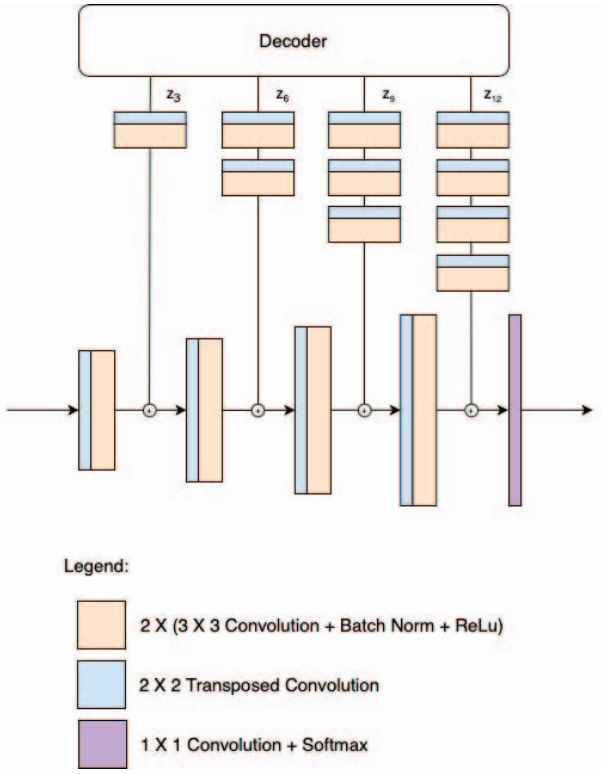

Fig. 2. Overall architecture of the feature aggregation branch. The convolutional layers from the volume-centric branch to the slice-centric branch are not drawn to scale.

$$L_{segmentation} = L_{dice} + L_{CE} \quad (2)$$

Hence, the final loss function used is the sum of the reconstruction loss and the segmentation loss:

$$L = L_{reconstruction} + L_{segmentation} \quad (3)$$

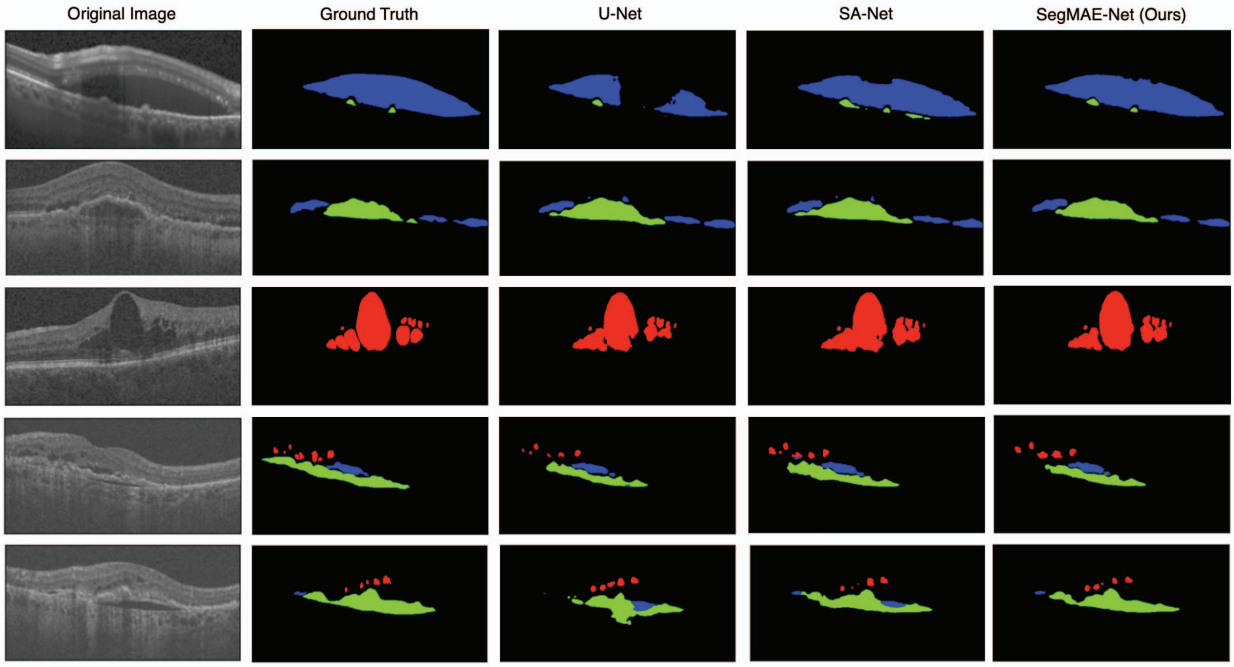

Fig. 3. Comparison of qualitative results between SegMAE-Net and other segmentation methods, where the red, green and blue regions represent the IRF, SRF and PED fluids respectively.

D. Feature Aggregation Branch

To fuse the spatial information of each slice and the contextual information learned from its adjacent slices, a feature aggregation branch connects the decoders of the volume-centric branch and the slice-centric branch. We extract the decoded feature representation $z_i \in \mathbb{R}^{\frac{HW}{P^2} \times D}, i \in \{3, 6, 9, 12\}$ and reshape it into a $\frac{H}{P} \times \frac{W}{P} \times D$ tensor, where D represents the dimension of the embedded space. This is followed by a series of 2×2 transposed convolutional layer and two 3×3 double convolutional layers, as shown in Fig. 2.

III. Experiments and Results

A. Dataset

We evaluate our proposed method on the Retinal OCT Fluid Challenge (RETOUCH) dataset [20]. This dataset consists of 70 optical coherence tomography (OCT) volumes, acquired with spectral domain SD-OCT devices from three different vendors, Cirrus, Spectralis and Topcon respectively. Each Cirrus and Topcon OCT consist of 128 B-scans, while each Spectralis OCT consist of 49 B-scans. Given the difference in sizes of B-scans across the three vendors, every B-scan is standardised to a size of 256×512 pixels after cropping out a region of interest based on the pixel intensity distribution. Three different retinal fluids were labelled, namely the intra-retinal fluid (IRF), sub-retinal fluid (SRF) and pigment epithelial detachment (PED). The actual test set for the RETOUCH dataset was unused in this study, given the lack of a ground truth segmentation to validate the results. Five-fold cross-validation was conducted with the training data provided, with 80% of OCT volumes acquired from each vendor as the training set and the remaining 20% to form our test set. We ensure an even representation of each vendor in our training and test sets during the cross-validation process.

B. Implementation Details

Our experiments were implemented on PyTorch, on a NVIDIA GeForce RTX 3090 GPU with 24GB of RAM. We initialise the weights of the network with He initialisation [21] and train the model for 20 epochs with an Adam optimiser [22]. We use a constant learning rate of 0.001 for the first ten epochs, followed by a learning rate of 0.0001 for the next ten epochs. We empirically use $n = 5$ as the number of adjacent slices to be stacked to form the volume input, and a batch size of 16. The convolutional filters used for the slice-centric branch are [16, 32, 64, 128, 256], increasing in the encoder path and decreasing in the decoder path. The filters used in the feature aggregation branch are [128, 64, 32, 16]. For the volume-centric path, a masking ratio of 75% is used. We use the Dice coefficient and Intersection over Union (IoU) to evaluate the segmentation accuracy in our experiments. As each volume is predicted slice-by-slice, the segmented 2D slices are then stacked together to form the 3D prediction for further evaluation. The two metrics used are defined as such

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

TABLE I

COMPARISON OF QUANTITATIVE RESULTS BETWEEN SEGMAE-NET AND OTHER SEGMENTATION METHODS (MEAN ± STANDARD DEVIATION) BY SLICE LEVEL. IRF, SFR AND PED REPRESENT INTRA-RETINAL FLUID, SUB-RETINAL FLUID AND PIGMENT EPITHELIAL DETACHMENT RESPECTIVELY. BEST RESULTS ARE MARKED IN BOLD.

Method	Dice (%)				IoU (%)			
	IRF	SRF	PED	Average	IRF	SRF	PED	Average
U-Net [8]	52.6 ± 6.6	**47.0 ± 9.3**	56.6 ± 12.8	52.1 ± 4.5	41.8 ± 5.5	38.2 ± 7.7	48.9 ± 10.8	43.0 ± 3.8
SA-Net [11]	52.8 ± 6.7	44.0 ± 11.1	59.1 ± 9.5	52.0 ± 4.3	42.2 ± 5.6	35.4 ± 9.8	51.2 ± 7.4	42.9 ± 3.6
SegMAE-Net (Ours)	**55.5 ± 5.9**	46.9 ± 10.2	**63.3 ± 7.3**	**55.2 ± 4.1**	**44.5 ± 5.0**	**38.3 ± 8.6**	**55.1 ± 5.7**	**46.0 ± 3.5**

TABLE II

COMPARISON OF QUANTITATIVE RESULTS BETWEEN SEGMAE-NET AND OTHER SEGMENTATION METHODS (MEAN ± STANDARD DEVIATION) BY VOLUME LEVEL. IRF, SFR AND PED REPRESENT INTRA-RETINAL FLUID, SUB-RETINAL FLUID AND PIGMENT EPITHELIAL DETACHMENT RESPECTIVELY. BEST RESULTS ARE MARKED IN BOLD.

Method	Dice (%)				IoU (%)			
	IRF	SRF	PED	Average	IRF	SRF	PED	Average
U-Net [8]	57.2 ± 7.4	40.3 ± 6.7	45.6 ± 13.4	47.7 ± 5.5	45.0 ± 5.4	32.6 ± 5.3	38.9 ± 12.1	38.8 ± 4.5
SA-Net [11]	56.4 ± 7.7	37.2 ± 8.6	49.9 ± 14.4	47.8 ± 4.7	45.1 ± 6.5	29.4 ± 6.8	42.3 ± 12.1	38.9 ± 4.0
SegMAE-Net (Ours)	**57.4 ± 5.5**	**42.4 ± 9.0**	**54.1 ± 9.9**	**51.3 ± 4.6**	**45.9 ± 4.7**	**34.2 ± 6.9**	**46.1 ± 8.0**	**42.1 ± 3.6**

$$IoU = \frac{TP}{TP + FP + FN} \tag{5}$$

where TP, FP and FN represent the number of true positive, false positive and false negative counts respectively.

C. Quantitative Evaluation and Comparison

To test the effectiveness of our proposed method, we compared it with previous state-of-the-art models for medical image segmentation, namely U-Net [8] and SA-Net [11]. For a fair comparison, we use the same parameters for the convolutional parts of each proposed method and evaluate the segmentation results based on each fluid present and the overall average performance across all fluid classes. The results of our evaluation in terms of slice level and volume level are reported in Tables I and II respectively. We observe that SegMAE-Net outperforms the other two methods in both slice and volume levels across all classes, with the only exception of sub-retinal fluid in the slice level, where SegMAE-Net is the second-best performing model. Importantly, SegMAE-Net performs significantly better than other proposed methods in the volume level, with a margin difference of more than 3% of the second-best performing model. Fig. 3 shows some qualitative results of the comparative studies between our proposed method and other methods.

To further emphasise the superiority of our proposed method, we show the improved consistency of segmentation across multiple test volumes in Fig. 4. We plot the Dice coefficients of all segmented slices from representative cases

TABLE III

EFFECT OF MASKING RATIO ON SEGMENTATION PERFORMANCE (MEAN ± STANDARD DEVIATION). BEST RESULTS ARE MARKED IN BOLD.

Masking Ratio	Slice Level		Volume Level	
	Dice (%)	IoU (%)	Dice (%)	IoU (%)
0.25	51.2 ± 6.8	42.7 ± 5.7	50.4 ± 7.0	41.6 ± 5.8
0.50	54.6 ± 3.0	45.3 ± 2.5	51.2 ± 7.2	42.1 ± 6.1
0.75	**55.2 ± 4.1**	**46.0 ± 3.5**	**51.3 ± 4.6**	**42.1 ± 3.6**

TABLE IV

EFFECT OF UPSAMPLING METHOD ON SEGMENTATION PERFORMANCE (MEAN ± STANDARD DEVIATION). BEST RESULTS ARE MARKED IN BOLD.

Upsampling Method	Slice Level		Volume Level	
	Dice (%)	IoU (%)	Dice (%)	IoU (%)
Bilinear Interpolation	54.4 ± 4.6	45.1 ± 3.7	50.6 ± 7.0	41.5 ± 5.7
Transposed Convolution	**55.2 ± 4.1**	**46.0 ± 3.5**	**51.3 ± 4.6**	**42.1 ± 3.6**

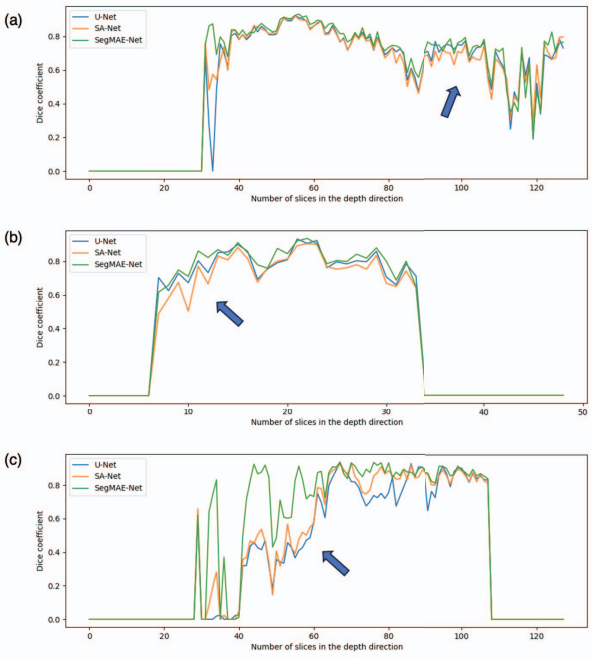

Fig. 4. Representative test volumes demonstrating the improved consistency by our proposed SegMAE-Net as compared to other segmentation methods. The graph plots the Dice coefficients against the position of each slice in the depth direction. (a), (b) and (c) represent IRF, SRF and PED respectively.

of IRF, SRF, and PED. These slices are arranged by their depth within the test volume. Notably, our SegMAE-Net is able to achieve a higher and more consistent performance in the depth direction. In contrast, the other two methods exhibit generally lower performance and more sudden fluctuations in the segmentation accuracy across the depth of the test volume. This highlights the potential of our method to be clinically valuable in situations where a 3D segmented structure is required to be extracted from a particular volume with high precision and consistency.

D. Ablation Study

From the comparison with previous existing methods, we prove that our volume-centric branch using MAE is a valuable component in capturing the volumetric context of the medical image to guide the segmentation. We also conduct experiments to study the effect of varying the masking ratio. As shown in Table III, a higher masking ratio results in better performance due to the ability of the MAE to acquire meaningful representations and learn the contextual information from the adjacent slices. We also compare the upsampling used to recover the segmentation output and prove that transposed convolutions are more effective in our model. The results are shown on Table IV.

IV. Conclusion

In this paper, we introduce SegMAE-Net, a hybrid 2D and 3D model for medical image segmentation. By using both masked autoencoders and convolutional neural networks, SegMAE-Net can extract global and local features respectively to improve volumetric segmentation. Experiments on the RETOUCH dataset have shown that our proposed method have outperformed previous state-of-the-art methods. We hope that our work can improve the speed of diagnosis by medical professionals through identifying, segmenting and analysing structures of interest in medical images effectively and efficiently. Particularly, we believe that our work can benefit doctors in various applications such as surgical planning, tumour monitoring and therapy optimisation.

Acknowledgement

This research received support from the Agency for Science, Technology and Research (A*STAR) AME Programmatic Funds (Grant Number : A20H4b0141).

References

[1] Y. Liu et al., "Disentangled representation learning for OCTA vessel segmentation with limited training data," *IEEE Trans. Med. Imaging*, vol. 41, no. 12, pp. 3686-3698, Dec. 2022, doi: 10.1109/TMI.2022.3193029.

[2] L. D. Le et al., "An efficient defending mechanism against image attacking on medical image segmentation models," in *Resource-Efficient Medical Image Analysis*, in Lecture Notes in Computer Science, vol. 13543, Sep. 2022, pp. 65-75, doi: 10.1007/978-3-031-16876-5_7.

[3] G. Tjio, P. Liu, J. T. Zhou, and R. S. M. Goh, "Adversarial semantic hallucination for domain generalized semantic segmentation," in *Proc. IEEE/CVF Winter Conf. Appl. Compute. Vision (WACV)*, 2022, pp. 318-327.

[4] V. Bellemo et al., "Cross-modality optical coherence tomography image enhancement using deep learning," *Investigative Ophthalmology & Visual Science*, vol. 64, no. 8, p. 235, Jun. 2023.

[5] S. Hussain et al., "Generative modelling for synthesis of cellular imaging data for low-cost drug repurposing application," in *Pacific-Asia Conf. Knowl. Discovery Data Mining*, 2020, pp. 165-177, doi: 10.1007/978-3-030-60470-7_16.

[6] Y. Zhao, W. Gui, Z. Chen, J. Tang, and L. Li, "Medical images edge detection based on mathematical morphology," in *2005 IEEE Eng. Medicine Biol. 27th Annu. Conf.*, Shanghai, China, 2006, pp. 6492–6495, doi: 10.1109/IEMBS.2005.1615986.

[7] C. Cigla and A. A. Alatan, "Region-based image segmentation via graph cuts," in *2008 15th IEEE Int. Conf. on Image Process.*, San Diego, CA, USA, 2008, pp. 2272-2275, doi: 10.1109/ICIP.2008.4712244.

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention*, in Lecture Notes in Computer Science, vol. 9351, Nov. 2015, pp. 234-241, doi: 10.1007/978-3-319-24574-4_28.

[9] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *Medical Image Computing and Computer-Assisted Intervention*, in Lecture Notes in Computer Science, vol. 9901, Oct. 2016, pp. 424-432, doi: 10.1007/978-3-319-46723-8_49.

[10] F. Milletari, N. Navab. and S. A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 4th Int. Conf. on 3D Vision (3DV)*, Stanford, CA, USA, 2016, pp. 565-571, doi: 10.1109/3DV.2016.79.

[11] D. A. Y. Cahyo et al., "Multi-task learning approach for volumetric segmentation and reconstruction in 3D OCT images," *Biomed. Opt. Express*, vol. 12, no. 12, pp. 7348-7360, Nov. 2021, doi: 10.1364/BOE.428140.

[12] H. Tang et al., "Spatial context-aware self-attention model for multi-organ segmentation," in *Proc. IEEE/CVF Winter Conf. Appl. Compute. Vision (WACV)*, 2021, pp. 939-949.

[13] A. Vaswani et al., "Attention is all you need," in *31st Conf. Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, 2017, pp. 5998-6008.

[14] A. Dosovitskiy et al., "An image is worth 16x16 words: transformers for image recognition at scale," Presented at ICLR 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy.

[15] A. Hatamizadeh et al., "UNETR: Transformers for 3D medical image segmentation," in *Proc. IEEE/CVF Winter Conf. Appl. Compute. Vision (WACV)*, 2022, pp. 574-584.

[16] J. Chen et al., "TransUNet: Transformers make strong encoders for medical image segmentation," 2021. [Online]. Available: arXiv: 2102.04306.

[17] S. Guo et al., "Causal knowledge fusion for 3D cross-modality cardiac image segmentation," *Inf. Fusion*, vol. 99, no. 101864, Nov. 2023, doi: 10.1016/j.inffus.2023.101864.

[18] K. He et al., "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Compute. Vision Pattern Recognit. (CVPR)*, 2022, pp. 16000-16009.

[19] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, pp. 448-456.

[20] H. Bogunović et al., "RETOUCH: The retinal OCT fluid detection and segmentation benchmark and challenge," in *IEEE Trans. Med. Imag.*, vol. 38, no. 8, pp. 1858-1874, Aug. 2019, doi: 10.1109/TMI.2019.2901398.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Compute. Vision*, 2015, pp. 1026-1034.

[22] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," 2014. [Online]. Available: arXiv: 1412.6980.