SUPERRESOLUTION AND SEGMENTATION OF OCT SCANS USING MULTI-STAGE ADVERSARIAL GUIDED ATTENTION TRAINING

Paria Jeihouni, Omid Dehzangi, Annahita Amireskandari, Ali Dabouei, Ali Rezai, Nasser M. Nasrabadi

Rockefeller Neuroscience Institute, Computer Science & Electrical Engineering, Ophthalmology & Visual Sciences, West Virginia University, , USA

{{*pj00001, ad0046*}*@mix*, {*omid.dehzangi@, annahita.amireskandari@, ali.rezai@*}*hsc, nasser.nasrabadi@mail*}*.wvu.edu*

ABSTRACT

Optical coherence tomography (OCT) is one of the non-invasive and easy-to-acquire biomarkers (the thickness of the retinal layers, which is detectable within OCT scans) being investigated to diagnose Alzheimer's disease (AD). This work aims to segment the OCT images automatically; however, it is a challenging task due to various issues such as the speckle noise, small target region, and unfavorable imaging conditions. In our previous work, we have proposed the multi-stage & multi-discriminatory generative adversarial network (MultiSDGAN) [1] to translate OCT scans in high-resolution segmentation labels. In this investigation, we aim to evaluate and compare various combinations of channel and spatial attention to the MultiSDGAN architecture to extract more powerful feature maps by capturing rich contextual relationships to improve segmentation performance. Moreover, we developed and evaluated a guided mutli-stage attention framework where we incorporated a guided attention mechanism by forcing an L-1 loss between a specifically designed binary mask and the generated attention maps. Our ablation study results on the WVU-OCT data-set in five-fold cross-validation (5-CV) suggest that the proposed MultiSDGAN with a serial attention module provides the most competitive performance, and guiding the spatial attention feature maps by binary masks further improves the performance in our proposed network. Comparing the baseline model with adding the guided-attention, our results demonstrated relative improvements of 21.44% and 19.45% on the Dice coefficient and SSIM, respectively.

Index Terms— Optical Coherence Tomography, Generative Adversarial Networks, Superresolution, MultiSDGAN, Attention Mechanism, Guided Attention

1. INTRODUCTION

AD is a progressive neurodegenerative disease that gradually declines memory and cognitive function. Previous studies have reported that the retina shares similar anatomical and physiological features with the brain, so it can be used as a possible biomarker for AD diagnosis in clinical practice [2]. Unlike current standard methods that are invasive and expensive for AD detection [3], the thickness of the retina layer can be noninvasively assessed using high-resolution images obtained with optical coherence tomography (OCT). Because of noise and artifacts (e.g., eye motions, the vessel projection shadow), manual segmentation of OCT images is a challenging task. Hence, it is imperative to program a method of OCT-based automatic retina layer segmentation.

Convolutional neural networks (CNNs) have achieved state-of-the-art performance in a breadth of image segmentation tasks, and they have robust and nonlinear feature extraction capabilities [4, 5]. Although U-Net [6] is a common network for medical image segmentation, it has issues dealing with class imbalance labels. The main problem is the usage of cross-entropy (CE) loss [7]. Since the foreground to background ratio is low in the medical images, using CE will learn a decision boundary biased towards the majority class, which would result in inaccurate segmentation. On the other hand, Generative Adversarial Networks (GANs) have been extensively used for various challenging medical segmentation tasks [8]. GANs have been quite prominent in learning deep representations and modeling high-dimensional data. Conditional GANs depict good performance translating data from one domain to another [9], [10], thus it is appropriate for semantic segmentation. In our previous works, we proposed a GAN-based domain translation and superresolution architecture that learns to increase the medical image resolution from low to high and learn to segment the retinal layers at the same time [11, 12, 1]. This particular type of GAN considers multiple stages of output from different layers of the network. Each intermediary output from the multi-stage is subjected to various discriminators [1].

Attention modules were widely used to boost segmentation performance [13]. The attention module allows the network to focus on the most relevant features without additional supervision, avoiding using multiple similar feature maps and highlighting salient features that are useful for a given task. Channel attention selects meaningful features at channel dimension, and spatial attention calculates the feature representation in each position by the weighted sum of the features from all the other positions [14]. Previous studies depict the importance of attention modules on improving the performance of OCT image segmentation [15]; however, to

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

the best of our knowledge, it is the first time that a GAN-based attention module has been employed for OCT segmentation.

In this paper, we aim to investigate the impact of incorporating attention mechanism in the MultiSDGAN framework [1]. We aim to evaluate and compare various combinations of channel and spatial attention to capture rich contextual relationships to extract more powerful feature maps and improve segmentation outcome. Therefore, we design a method to train the network based on multi-stage attention modules and assess if there is any improve in segmentation performance. Moreover, this study take a step further and add regularization on top of the spatial attention feature maps to focus the attention on our region of interest, which we refer to that as the guided attention module. In this way, the guided attention modules are added to the generator by forcing the L-1 loss between a specifically designed binary mask and the attention maps generated at different layers of the network and we investigate its effectiveness in improving the final results. To the best of our knowledge, it is the first time that this study has been investigated. The rest of this paper is focused on introducing and evaluating the impact of various combinations of channel and spatial attention in multi-stages with self and guided attention approach on the MultiSDGAN model.

2. DATA ACQUISITION AND PREPROCESSING

Participants are recruited based on referrals to current patients at the memory disorders clinic or geriatric clinic at the West Virginia University (WVU). An ophthalmologist conducted a complete eye exam on all the subjects, including visual acuity, intraocular pressure, pupillary reaction, and dilated fundus exam. The Heidelberg Spectralis OCT (Heidelberg Engineering Inc., Heidelberg, Germany) was used to obtain the OCT of the macula and the optic nerve head.

Data collection was initiated with normal aging patients (age: 55+). The Ophthalmology Department at the WVU medicine provided the OCT images of 55 subjects, each having 19 scans and six subjects had one extra OCT. In total, our data-set has 1,045 images. These are 2-D scans, each group of 19 constitute one 3-D scan of the macula. Each image was meticulously labeled each image for the 7 innermost layers by an expert in the field. Finally, all patient data was de-identified prior to analysis based on the WVU Institutional Review Board (IRB) approved under the study ID: 1910761036. Horizontal flipping, spatial translation, and rotation are among the methods we employed to augment the data-set[16]. Also, to increase the data-set size synthetically, we used a moving crop window approach of size 224x224, which was moves on the image with 75% overlap.

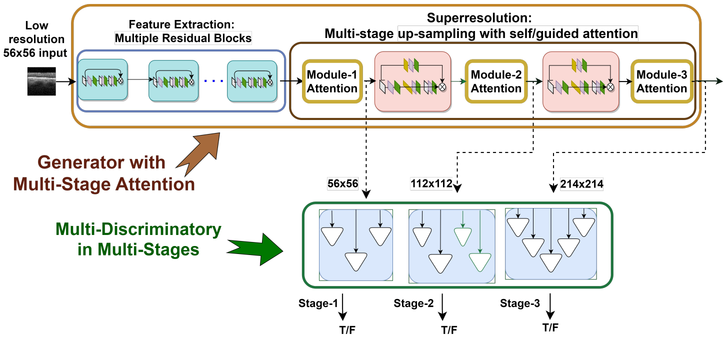

Low resolution 56x56 input
Feature Extraction: Multiple Residual Blocks
Superresolution: Multi-stage up-sampling with self/guided attention
Module-1 Attention
Module-2 Attention
Module-3 Attention
Generator with Multi-Stage Attention
56x56
112x112
214x214
Multi-Discriminatory in Multi-Stages
Stage-1
T/F
Stage-2
T/F
Stage-3
T/F

Fig. 1: Our Multi-Stage Multi-Discriminatory GAN (MultiSDGAN) architecture is used as the framework for superresolution and segmentation. Multi-stage generator G consists of several residual blocks and transposed blocks with added attention mechanisms in various stages of superresolution. The multi-discriminatory modules provide scrutiny at different patch levels.

3. METHODOLOGY

3.1. MultiSDGAN

Generative adversarial network contains of two subnetworks: the generator and the discriminator. Fig.1 illustrates our GAN-based domain translation framework, MultiSDGAN. MultiSDGAN adopts and modifies ResNet as its generator architecture. The generator that is employed in this architecture has two major parts. The first part is being used for extracting features, and the second part superresolves the images to a certain scale. To achieve the superresolution, a transposed bottleneck block was designed to be added to the generator. One important feature of this generator is its multi-stage output, which is basically extracting outputs from different intermediatory layers of the network, rather than only the final layer as suggested in [17]. Additionally, multiple discriminators are being used to enhance the discriminatory aspects of the GAN. Each of these discriminators is a PatchGAN [9], in which a convolutional neural network classifies an image as fake or real by focusing on penalizing it at the scale of local image patches of size $N \times N$.

3.2. Attention module

Attention mechanisms allow humans to selectively focus on key information while ignoring other irrelevant information. Through the attention module, deep CNN can extract more critical and discriminative features for the target task, and enhance the robustness of the network model [18].

3.2.1. Channel attention module

The channel attention module is used to selectively weight the importance of each channel and thus produces best output features. This helps in reducing the number of parameters of

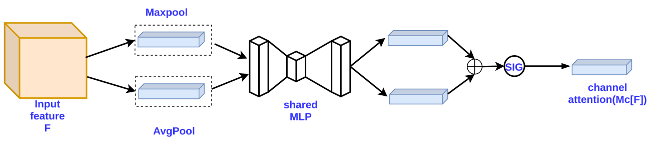


Fig. 2: Channel attention module.

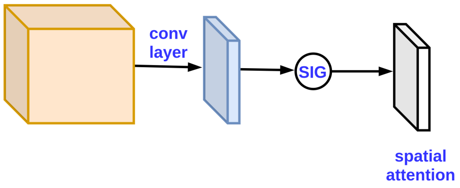


Fig. 3: Spatial attention module.

the network. To compute the channel attention, the spatial dimension of the input feature map was squeezed by average-pooling and Max-pooling [19]. Fig. 2 illustrates the channel attention mechansim. In short, the channel attention is computed as:

$$\mathrm{M}_c(F) = \sigma(MLP[AvgPool(\mathrm{F})]) + \sigma(MLP[MaxPool(\mathrm{F})]). \quad (1)$$

3.2.2. Spatial attention module

This module is designed to learn the spatial dependencies in the feature maps. Specifically, a depth-wise convolution is used to extract information to have distant vision over the feature maps. To compute spatial attention, we apply 1×1 convolution instead of max-pooling and avg-pooling to decrease the depth of the feature maps. In this way, the model learns how to shrink the depth by keeping the most relevant information. Fig. 3 illustrates the spatial attention mechanism. In short, the spatial attention is computed as:

$$M_s(F) = \sigma(conv_{1x1}(\mathrm{F})). \quad (2)$$

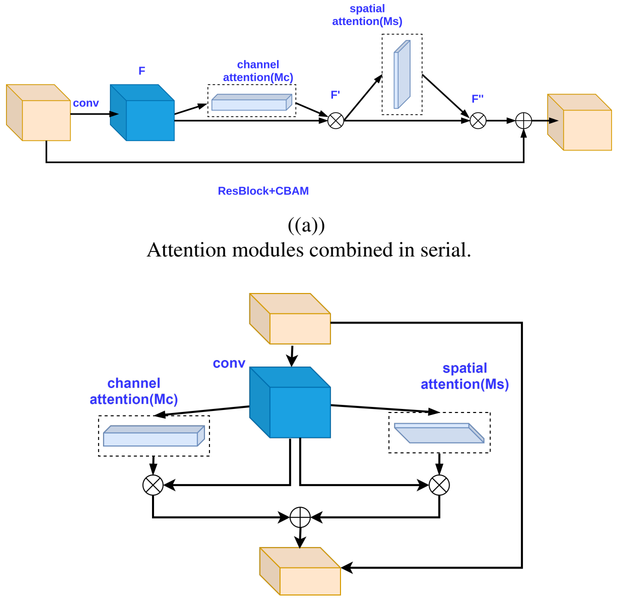
As we discussed previously, we want to investigate the impact of the serial and the parallel attention module beside individual channel and spatial attention. Fig. 4(a) and 4(b) depicts the architecture of sequential and parallel attention.

3.2.3. Guided Attention Module

As discussed earlier, positional information of images is the main focus of spatial attention, and it can detect the spatial relationship between the input features [20]. To improve the performance of spatial attention even further, we adopt a binary mask to guide the feature maps for spotlighting the region of interest. Fig.5(b) depicts the binary mask in which only the region of interest (i.e., inner retina layers) is white.

3.3. Loss function

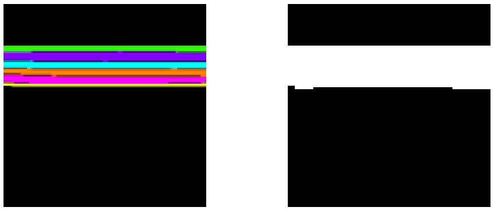
Similar to the MultiSDGAN model [1], during the training, the proposed network weights are updated based on the Dice, SSIM and L-1 losses that are obtained via the following formulas. y is our ground truth and G(x) is the generator output.



((a))
Attention modules combined in serial.

((b))
Attention modules combined in parallel.

Fig. 4: Sequential and parallel combinations of channel and spatial attention modules.



((a))
label.

((b))
Binary Guided attention mask.

Fig. 5: The attention is guided to by focus on the region of interest shown in (a) by enforcing the attention mask in (b).

$$L_{L-1}(G) = \|\mathrm{y} - G(\mathrm{x})\|_1, \quad (3)$$

$$L_{Dice}(G) = 1 - \frac{2\sum_{i=1}^{n} G(x_i) y_i}{\sum_{i=1}^{n} G(x_i)^2 + \sum_{i=1}^{n} y_i^2}, \quad (4)$$

$$SSIM(G(\mathrm{x}), \mathrm{y}) = \frac{(2\mu_{G(\mathrm{x})}\mu_{\mathrm{y}} + c_1)(2\sigma_{G(\mathrm{x}),\mathrm{y}} + c_2)}{(\mu_{G(\mathrm{x})}^2 + \mu_{\mathrm{y}}^2 + c_1)(\sigma_{G(\mathrm{x})}^2 + \sigma_{\mathrm{y}}^2 + c_2)}. \quad (5)$$

4. EXPERIMENTS & RESULTS

In this section, the proposed architecture will be analyzed using various combinations of attention modules. Then, the whole architecture including applying attention modules on

Table 1: SSIM, Dice coef and L-1 comparison among various combinations of attention modules.

Model			Dice Coeff.			SSIM			L-1		
Measure			Last Stage Attention	Multi-Stage Attention	No Attention	Last Stage Attention	Multi-Stage Attention	No Attention	Last Stage Attention	Multi-Stage Attention	No Attention
MultiSDGAN with attention mechanism	Channel		0.9076±0.007	0.9102 ±0.008	0.9016±0.004	0.9012±0.0048	0.9086±0.0038	0.8987±0.0051	0.020 ±0.0012	0.019 ±0.0012	0.021 ±0.0011
	Spatial		0.9105 ±0.005	0.9076 ±0.008	0.9016±0.004	0.9038±0.0022	0.9093±0.0034	0.8987±0.0051	0.020±0.0011	0.018±0.0012	0.021±0.0011
	Ch&Sp	Parallel	0.9132±0.006	0.9134±0.002	0.9016±0.004	0.9087±0.0024	0.9145±0.0036	0.8987±0.0051	0.018±0.0012	0.017±0.0015	0.021±0.0011
		Sequential	0.9158±0.002	**0.9187±0.0011**	0.9016±0.004	0.9125±0.0034	**0.9153±0.0024**	0.8987±0.0051	0.018±0.0011	**0.016±0.0010**	0.021±0.0011

multiple stages will be discussed, and the results will be compared. In the end, the effect of using attention mask will be explained. The parameters set in the network are as follows: Learning rate= 0.001, loss function= *Dice, SSIM, L-1*, optimizer function= *Adam*, batch size= 8, number of epochs= 200. Furthermore, we have divided the data-set into two parts, the train set and the validation set, where the percentage of the divisions are 80% and 20% (5-CV), respectively.

4.1. Impact of self-attention mechanisms

The impact of four attention modules, namely, single spatial attention, single-channel attention, parallel attention, and Sequential attention, are investigated in this study. This comparison aims to find the best attention module for the ask in hand. In the first stage, we applied these attention modules to the last layer. As shown in Table 1, the best-achieved performance belongs to the Sequential attention module consistently on the Dice coefficient, SSIM, and L-1 loss. The next best result is achieved by the parallel attention module.

4.2. Impact of multi-stage self-attention mechanisms

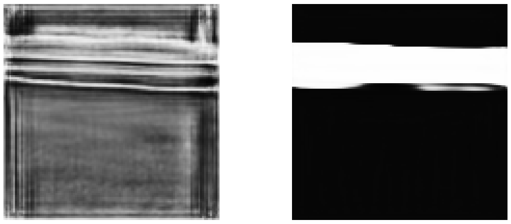
In the next stage, we applied the attention modules to all stages of suerresolution in the MultiSDGAN framework. As shown in Table 1, the multi-stage extension (see Fig. 1) improves the evaluation criteria consistently. The best-achieved performance belongs to the sequential attention module in multi-stages and the parallel attention is the next best.

4.3. Impact of multi-stage guided-attention mechanisms

The reported results in Table 1 depict that we get the best performance from the Sequential attention module. We further improved the results via guiding the training using a binary mask. As it can be seen in Table 2, applying binary mask further improved the performance of Sequential attention in comparison with the baseline without any attention mechanism. It demonstrated relative improvements of 21.44% (p_value<0.05, t-test on mean differences) and 19.45% (p_value<0.05) on the Dice coefficient and SSIM, respectively. Also, Table 2 demonstrate that all variants of our proposed adversarial attention mechanisms provide improved results in comparison with RelayNet [16], as a strong baseline model (with the impressive highest relative improvement of 41.16% and p_value<0.01).

Table 2: SSIM, Dice coef and L-1 comparison among MultiSDGAN, MultiSDGAN with Sequential attention module and MultiSDGAN with Sequential guided attention module.

Model		Dice Coefficient	SSIM	L-1
RelayNet		0.8828±0.0017	0.8613±0.0023	0.027±0.0011
MultiSDGAN	**No Attention**	0.9016±0.003	0.8987±0.0051	0.021±0.0011
	Self-attention	0.9187±0.0011	0.9153±0.0024	0.016±0.0010
	Guided-attention	**0.9227±0.0022**	**0.9184±0.0054**	**0.016±0.0009**



((a))
Trained self-attention feature map.

((b))
Trained guided-attention feature map.

Fig. 6: The final trained attention feature maps for the cases of (a) self- and (b) guided-attention.

5. CONCLUSION

In this paper, we proposed a new feature on our MultiSDGAN segmentation network that can refine features based on various attention modules. Experiments on our own data-set demonstrated the effectiveness of the attention mechanisms on MultiSDGAN. Sequential combination and guided attention mechanism provided the best empirical results by reducing the redundancy in model training. We aim to design attention modules effectively and capture more discriminative features for semantic inference as our future direction.

6. REFERENCES

[1] P. Jeihouni, O. Dehzangi, A. Amireskandari, A. R. Rezai, and N. M. Nasrabadi, "Multisdgan: translation of oct images to superresolved segmentation labels using multi-discriminators in multi-stages," *IEEE Journal of Biomedical and Health Informatics*, 2021.

[2] D. Sánchez, M. Castilla-Marti, M. Marquié, S. Valero, S. Moreno-Grau, O. Rodríguez-Gómez, A. Piferrer, G. Martínez, J. Martínez, I. De Rojas *et al.*, "Evaluation of macular thickness and volume tested by optical coherence tomography as biomarkers for alzheimer's disease in a memory clinic," *Scientific reports*, vol. 10, no. 1, pp. 1–9, 2020.

[3] L. K. Ferreira and G. F. Busatto, "Neuroimaging in alzheimer's disease: current role in clinical practice and potential future applications," *Clinics*, vol. 66, pp. 19–24, 2011.

[4] A. Sinha and J. Dolz, "Multi-scale self-guided attention for medical image segmentation," *IEEE journal of biomedical and health informatics*, vol. 25, no. 1, pp. 121–130, 2020.

[5] O. Dehzangi, P. Jeihouni, V. Finomore, and A. Rezai, "Physiological monitoring of front-line caregivers for cv-19 symptoms: Multi-resolution analysis amp; convolutional-recurrent networks," in *2021 IEEE International Conference on Image Processing (ICIP)*, 2021, pp. 250–254.

[6] N. Siddique, P. Sidike, C. Elkin, and V. Devabhaktuni, "U-net and its variants for medical image segmentation: theory and applications," *arXiv preprint arXiv:2011.01118*, 2020.

[7] B. Murugesan, K. Sarveswaran, S. M. Shankaranarayana, K. Ram, M. Sivaprakasam *et al.*, "A context based deep learning approach for unbalanced medical image segmentation," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2020, pp. 1949–1953.

[8] B. Lei, Z. Xia, F. Jiang, X. Jiang, Z. Ge, Y. Xu, J. Qin, S. Chen, T. Wang, and S. Wang, "Skin lesion segmentation via generative adversarial networks with dual discriminators," *Medical Image Analysis*, vol. 64, p. 101716, 2020.

[9] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[10] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[11] P. Jeihouni, O. Dehzangi, A. Amireskandari, A. Rezai, and N. M. Nasrabadi, "Gan-based super-resolution and segmentation of retinal layers in optical coherence tomography scans," in *2021 IEEE International Conference on Image Processing (ICIP)*, 2021, pp. 46–50.

[12] O. Dehzangi, S. H. Gheshlaghi, A. Amireskandari, N. M. Nasrabadi, and A. Rezai, "Oct image segmentation using neural architecture search and srgan," in *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 6425–6430.

[13] Y. Liu, Y. Chen, P. Lasang, and Q. Sun, "Covariance attention for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[14] A. Sagar, "Dmsanet: Dual multi scale attention network," *arXiv preprint arXiv:2106.08382*, 2021.

[15] D. Li, M. Zhang, W. Shi, H. Zhang, D. Wang, and L. Wang, "Pyramid pooling channel attention network for esophageal tissue segmentation on oct images," in *2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*. IEEE, 2020, pp. 1476–1480.

[16] A. G. Roy, S. Conjeti, S. P. K. Karri, D. Sheet, A. Katouzian, C. Wachinger, and N. Navab, "Relaynet: retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional networks," *Biomedical optics express*, vol. 8, no. 8, pp. 3627–3642, 2017.

[17] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "Stackgan++: Realistic image synthesis with stacked generative adversarial networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 8, pp. 1947–1962, 2018.

[18] X. Tong, J. Wei, B. Sun, S. Su, Z. Zuo, and P. Wu, "Ascunet: Attention gate, spatial and channel attention u-net for skin lesion segmentation," *Diagnostics*, vol. 11, no. 3, p. 501, 2021.

[19] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[20] Y. Lee and J. Park, "Centermask: Real-time anchor-free instance segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 13 906–13 915.