COUPLED SWIN TRANSFORMERS AND MULTI-APERTURES NETWORK(CSTA-NET) IMPROVES MEDICAL IMAGE SEGMENTATION

Siyavash Shabani, Muhammad Sohaib, Sahar A Mohamed *Student Member*, *IEEE*, Bahram Parvin, *Senior Member, IEEE*
Department of Electrical and Biomedical Engineering, University of Nevada, Reno

2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI) | 979-8-3315-2052-6/25/$31.00 ©2025 IEEE | DOI: 10.1109/ISBI60581.2025.10981294

ABSTRACT

Vision Transformers have outperformed traditional convolution-based frameworks across various visual tasks, including, but not limited to, the segmentation of 3D medical images. To further advance this area, this study introduces the Coupled Swin Transformers and Multi-Apertures Networks (CSTA-Net), which integrates the outputs of each Swin Transformer with an Aperture Network. Each aperture network consists of a convolution and a fusion block for combining global and local feature maps. The proposed model has been tested on two independent datasets to show that fine details are delineated. The proposed architecture was trained on the Synapse multi-organ and ACDC datasets to conclude an average Dice score of 90.19±0.05 and 93.77±0.04, respectively. The code is available here: https://github.com/Siyavashshabani/CSTANet.

Index Terms— Medical Image Segmentation, Vision Transformer, Distance Transform

1. INTRODUCTION

Recently, Vision Transformers (ViT) [4] have demonstrated state-of-the-art performance on many image classification benchmarks[6]. They have been applied in medical image segmentation [11, 12]. H. Cao and colleagues [11] introduced a novel Transformer-based architecture that replaces the traditional UNet framework for medical image segmentation. Because of the added value of the convolutional blocks, there has been a growing interest in developing methods to integrate representations from the Transformers and convolutional modules to preserve local information. Among these, Swin UNETR [5, 15] employs the Swin technique to construct a UNet-shaped architecture featuring a Transformer-based encoder and a convolutional-based decoder.

In support of the proposed research, it is noteworthy to indicate that the UNet model and its extensions have been examined extensively for biological applications [16, 17]. In basic UNet-CNN models[2, 18] , the encoder block maps the image to a lower-dimensional (latent) space, then reconstructed by the decoder block. Additionally, the skip connections at each dimensionality reduction stage map the encoder output to the decoder blocks to improve medical image segmentation [19]. These blocks excel in extracting local image features but are challenged to extract global features and their associations. Some studies have incorporated the attention mechanism [20-22] to enhance performance by improving the extraction of global features. Although recent advances can grasp local and global contexts, we hypothesized that improved 3D segmentation could benefit from multi-aperture representation for a more accurate representation of surfaces. In this study, we proposed a new architecture based on coupled Swin Transformers and Aperture blocks for capturing multi-aperture representations. Here, multi-aperture refers to overlapping multiscale representations of a region, where each aperture maintains its original resolution. Hence, the surfaces between adjacent objects are not diffused, as is the case for multiscale representation. The proposed architecture has been applied to two datasets of the Synapse [23] and ACDC [24] datasets.

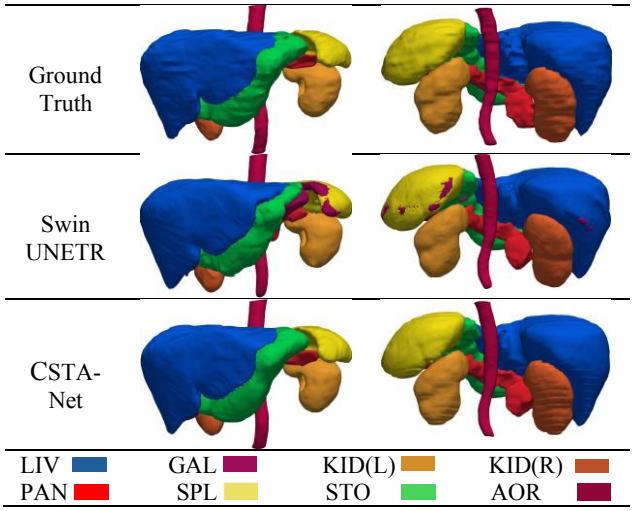

Fig. 1. CSTA-Net architecture shows an improved qualitative behavior when compared to the prior art. Two independent samples are shown per column.

The main contributions of our study are (i) four coupled Swin Transformers and aperture modules that reconstruct the pyramid representation with the original resolution, (ii) a 3D fusion block with a squeeze and excitation block and a convolutional block attention module (CBAM) for combining the outputs of each Transformer and its

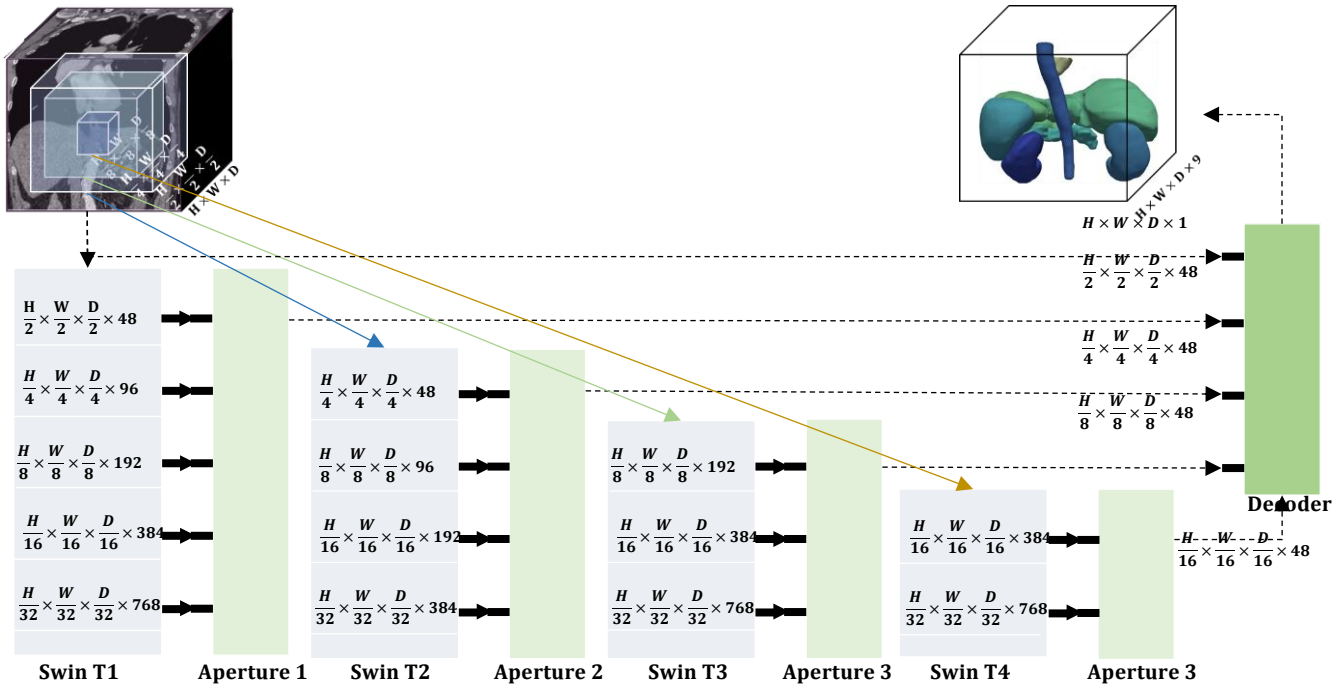

Fig. 2. CSTA-Net framework reconstructs the output mask via a decoder, whose inputs are created by coupled Swin Transformers and Aperture blocks in multi-scale representations.

convolutional block, and (iii) a loss function that integrates the Dice cross-entropy with a custom distance transform for morphometric consistency.

In summary, we (a) present improvements in delineating surface boundaries using the coupled Swin Transformers and Aperture blocks and (b) propose a 3D fusion block for integrating the global and local features that are outputs of Transformers and convolutional blocks, respectively.

The structure of this paper is organized as follows: Section 2 offers a concise review of the existing literature. Section 3 details the methodology used and its implementation. Section 4 presents the findings, includes visualizations and an ablation study, and provides the study's conclusion.

2. RELATED WORKS

Initial Vision Transformers (ViTs [4]) implementations were developed for image classification. Vanilla ViT employs a self-attention mechanism that utilizes pairwise similarities of input patches to capture long-range dependencies, enhancing performance relative to traditional architectures such as ResNet [6] or the basic UNet [16]. Recent studies [25-27] have shown that combining the transformer-based and convolutional-based modules enhances the performance of medical segmentation tasks. For example, 3D TransUNet [14] integrates the Transformer blocks to capture long-range dependencies in the bottleneck of UNet-shape architecture, which has lower dimension space, to enhance performance. Zhou et al. [7] introduced a novel architecture incorporating local self-attention modules within the encoder and decoder blocks, where global self-attention modules are integrated at the bottleneck of their framework. Other studies, such as UNETR++ [10] and Swin UNETR [15], primarily focus on researching combining convolutional operations and Transformer architectures to enhance model performance.

3. METHOD

This section presents the proposed CSTA-Net, its implementation, and the customized loss function. Figure 2 presents the proposed architecture, composed of four interconnected Swin Transformers (e.g., Swin T1, Swin T2, Swin T3, Swin T4) and Aperture blocks dedicated to multi-aperture reconstruction. The outputs of all paired sets are then passed to a decoder block that reconstructs the output masks. The Decoder block, borrowed from [15], takes its input from the Aperture blocks and upsamples to reconstruct the proper instance map. Similar approaches have been used in the UNet or Swin UNETR decoder block. This section provides further information on the fusion, Aperture blocks, loss function, datasets, and implementation details.

3.1. Fusion Block

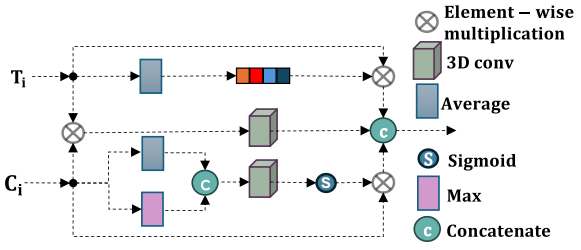

Fig. 3. The fusion block.

Fusion blocks are designed for integrating the global and local spatial features, shown in Fig. 3 Each fusion block

concatenates the outputs of the (i) Transformer following the squeeze and excitation (SE) block [28] (color-coded block), (ii) element-wise multiplication of the output of the Transformer and Convolution, and (iii) convolutional module following the Convolutional Block Attention Module (CBAM) [29]. Each of these outputs accentuates a specific aspect of the outputs of the Swin Transformers. For example, the SE block prioritizes the significance of channel-wise information, simple element-wise multiplication (e.g., Hadamard dot product) modulates the importance of features, and CBAM weights spatial similarities.

3.2. Aperture Blocks

As presented in Figure 2, our CSTA-Net framework comprises four Swin Transformer blocks, each coupled with the Aperture block. In more detail, each paired Swin Transformer and Aperture is tasked with reconstructing a specific representation of the multi-aperture for the 3D input patch [H, W, D]. As shown in Fig. 4, the structure of each Aperture block consists of reconstruction layers at varying representation scales. The 3D convolution and fusion blocks are incorporated in each layer, followed by up-sampling blocks. The output of the up-sampling blocks, similar to the Vanilla U-Net structure, is passed to the next layer.

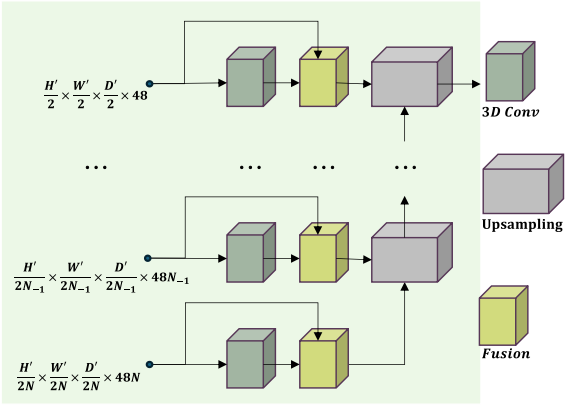

Fig. 4. Aperture block contains n branches based on inputs which are from their coupled Swin Transformers.

3.3. Loss Function

The proposed loss function is computed by the combined errors computed from the DiceCE and the distance transform map. For each class $_i$, let D_i be the absolute value of the signed distance transform of the predictor and let S_i be the indicator function on the ground truth, set to one at the surface points and zero elsewhere. The distance transform is normalized between 0 and 1. With λ as a hyperparameter, the combined loss function is defined as:

$$Loss(G,P) = DiceCE + \lambda.\frac{\sum D_i.S_i}{\sum D_i} \quad (1)$$

The goal is to measure the loss as a deviation from the ground truth. With the grid points at each surface location set to an indicator value of 1 in the ground truth, the loss function measures the closest surface point (e.g., distance transform) from the ground truth to the predictor. As a result, surfaces can be better delineated to match the ground truth.

3.4. Datasets

In this study, for the evaluation of our proposed framework, we used two public datasets: Synapse[1] multi-organ segmentation and ACDC[2] Datasets.

3.5. Implementation Details

Experiments were conducted on a Linux Cluster Server with 8 NVIDIA RTX 3080 GPUs, each with 12 GB of memory. We followed the method described in [15]; all images of the Synapse and ACDC datasets were resized with the exact physical dimensions of 1×1×1mm; subsequently, 128×128×128 random patches were extracted to build a dataset with the augmentation techniques outlined in [5]. The model was trained for 300 epochs for patches randomly selected from training images, and during the training process, the model was evaluated using validation images. Finally, evaluation was performed on test images, and the best-performing model within the 300 epochs was saved. Learning is based on Adam optimization with a learning rate of 10^{-4}, and a weight decay of 5×10^{-5}.

4. RESULTS

We evaluated and compared the proposed architecture with state-of-the-art (SOTA) in the field of medical image segmentation on the Synapse and ACDC datasets.

4.1. Performance on Synapse Dataset

Table 1 shows the comparative performance of the proposed model against prior research on the Synapse multi-organ dataset, where CSTA-Net produces a mean Dice score of 90.19±0.05 and HD95 7.12±0.04 with five-fold cross-validation. On average, these scores are better than the state-of-the-art studies 3D TransUNet and UNETR++. Moreover, out of eight classes of organs, we observed notable improvement in four: Kid(R), Kid(L), Aor, and Pan. Fig. 1 provides a comparative visualization of the 3D segmentation results of two image samples with Swin UNETR. This result illustrates reduced spurious regions and continuity within each organ, which are visible for Swin UNETR and nnFormer.

4.2. Performance on ACDC Dataset

Table 2 compares the performance of CSTA-Net architecture against nnFormer, UNETR++, and MIST with five-fold cross-validation as an average Dice 93.77±0.04. Accordingly,

[1] https://www.synapse.org/#!Synapse:syn3193805/wiki/217789
[2] https://www.creatis.insa-lyon.fr/Challenge/acdc/

Methods	Avg Dice↑	Avg HD95 ↓	Spl	Kid(R)	Kid(L)	Gal	Liv	Sto	Aor	Pan
nnU-Net [2]	87.33	-	91.68	88.46	84.68	78.82	97.13	83.34	93.04	81.50
MIST [3]	86.92	11.07	92.83	93.28	92.54	74.58	94.94	**87.23**	89.15	72.43
Swin UNETR [5]	83.47	10.55	95.37	86.26	86.99	66.54	95.72	77.01	91.12	68.80
nnFormer [7]	86.57	10.63	90.51	86.25	86.57	70.17	96.84	86.83	92.04	83.35
UNETR++ [10]	87.22	07.53	**95.77**	87.18	87.54	71.25	96.42	86.01	92.52	81.10
MISSFormer [13]	81.96	18.20	91.92	82.00	85.21	68.65	94.41	80.81	86.99	65.67
LeVit-UNet [8]	78.53	16.84	88.86	80.25	84.61	62.23	93.11	72.76	87.33	59.07
3DTransUNet [14]*	88.10	-	93.39	87.47	85.76	**81.15**	**97.34**	85.31	92.97	81.76
CSTA-Net	**90.19**	**07.12**	95.47	**93.56**	**93.70**	78.93	96.64	86.38	**93.13**	**83.72**

Table 1. CSTA-Net architecture improves the average Dice score on the synapse dataset over the prior art. Evaluation is performed against multiple organs (Gal for Gallbladder, kid(L) for Left Kidney, Kid(R) for Right Kidney, Pan for Pancreas, Spl for Spleen, Sto for Stomach, Liv for Liver, and Aor for Aorta) and reported as the mean Dice scores. The highest and second scores are emphasized in bold and underlined, respectively. the results are reported following five-fold cross-validation. *The HD95 for 3DTransunet has not been reported.

CSTA-Net not only has a better average Dice score than other methods, but it also performs better on each organ. Further visualization of the segmentation results on two 2D slices, as shown in Fig. 3, indicates that CSTA-Net is closer to the ground truth and generates no fragmentation.

Methods	Avg. Dice	RV	Myo	LV
VIT-CUP [1]	81.45	81.46	70.71	92.18
R50-VIT-CUP [1]	87.57	86.07	81.88	94.75
Swin UNETR [5]	88.61	85.29	86.52	94.02
LeViT-UNet [8]	90.32	89.55	87.64	93.76
SegFormer3D [9]	90.96	88.50	88.86	95.53
nnFormer [7]	92.06	90.94	89.58	95.65
MIST [3]	92.56	91.23	90.31	96.14
UNETR++ [10]	92.83	91.89	90.61	96.00
CSTA-Net	**93.77**	**92.94**	**91.85**	**96.53**

Table 2. CSTA-Net has a better average Dice score than SOTA.

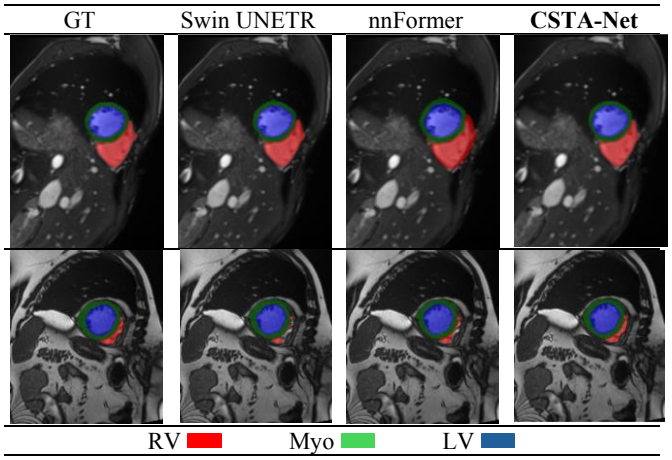

Fig. 3. CSTA-Net compares better qualitatively with the ground truth without fragmentation.

5. ABLATION STUDY

Effect of alternative loss functions: We examined the proposed framework using various objective function combinations, as detailed in Table 3. Introducing a novel term, DistLoss, into the mix with Dice loss and cross-entropy loss significantly enhanced performance on Synapse and ACDC datasets, culminating in the highest average Dice scores of 90.19 and 93.77, respectively.

Loss Function	Avg. Dice Synapse	Avg. Dice ACDC
L_{dice}	89.21	92.57
L_{focal}	89.60	92.38
$L_{dice} + L_{ce}$	89.43	92.91
$L_{dice} + L_{ce} + DistLoss$	**90.19**	**93.77**

Table 3. The ablation study indicates an improved performance with the inclusion of DistLoss function.

Blocks	Avg. Dice Synapse	Avg. Dice ACDC
$CSTA_1$	87.68	90.76
$CSTA_1 + CSTA_2$	88.10	91.44
$CSTA_1 + CSTA_2 + CSTA_3$	89.28	92.39
$CSTA_1 + \cdots + CSTA_4$	**90.19**	**93.77**

Table 4. The ablation study indicates the effectiveness of alternative combinations of modules on the performance of CSTA-Net.

Effect of Coupled Swin Transformer and Aperture Blocks: Table 4 summarizes the impact of our proposed coupled Swin Transformer and Aperture blocks on model performance for Synapse and ACDC datasets.

6. DISCUSSION

We proposed a new framework (CSTA-Net), that includes four coupled Swin Transformers and new Aperture Blocks for multi-scale representations for medical image segmentation. The model was evaluated using the Synapse and ACDC datasets. In our CSTA-Net, initially, the first coupled Transformer and Aperture Block receive an entire patch, while others receive scaled-down patches. As a result, the details are better preserved by maintaining the original image resolution. Although the multi-aperture approach increases the computational load, this increase is modest since the patch sizes are reduced exponentially. We also introduced a 3D fusion block for integrating the outputs of Transformers and convolution modules. Finally, we showed that a loss function enhances performance. Our model outperformed published research on Synapse and ACDC datasets, achieving a new state-of-the-art regarding the mean Dice and mean HD95 scores.

Acknowledgment: This research was supported by a grant from NIH CA279408.

REFERENCES

[1] J. Chen *et al.*, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306,* 2021.

[2] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods,* vol. 18, no. 2, pp. 203-211, 2021.

[3] M. M. Rahman, et al., "MIST: Medical Image Segmentation Transformer with Convolutional Attention Mixing (CAM) Decoder," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024.

[4] A. Dosovitskiy, et al, "An image is worth 16x16 words: Transformers for image recognition at scale.," *arXiv preprint,* 2020.

[5] A. Hatamizadeh, et al., "Unetr: Transformers for 3d medical image segmentation.," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision.*, 2022.

[6] K. He, et al., "Deep residual learning for image recognition," *Proceedings of the IEEE conference on computer vision and pattern recognition,* 2016.

[7] H.-Y. Z. e. al., "nnFormer: Volumetric Medical Image Segmentation via a 3D Transformer," *IEEE Transactions on Image Processing,* pp. 4036-4045, 2023.

[8] G. Xu, X. Zhang, X. He, and X. Wu, "Levit-unet: Make faster encoders with transformer for medical image segmentation," in *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, 2023: Springer, pp. 42-53.

[9] S. Perera, P. Navard, and A. Yilmaz, "SegFormer3D: an Efficient Transformer for 3D Medical Image Segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4981-4988.

[10] A. M. Shaker, M. Maaz, H. Rasheed, S. Khan, M.-H. Yang, and F. S. Khan, "UNETR++: delving into efficient and accurate 3D medical image segmentation," *IEEE Transactions on Medical Imaging,* 2024.

[11] H. Cao *et al.*, "Swin-unet: Unet-like pure transformer for medical image segmentation," in *European conference on computer vision*, 2022: Springer, pp. 205-218.

[12] A. Lin, B. Chen, J. Xu, Z. Zhang, G. Lu, and D. Zhang, "Ds-transunet: Dual swin transformer u-net for medical image segmentation," *IEEE Transactions on Instrumentation and Measurement,* vol. 71, pp. 1-15, 2022.

[13] X. Huang, Z. Deng, D. Li, X. Yuan, and Y. Fu, "MISSFormer: an effective transformer for 2D medical image segmentation," *IEEE transactions on medical imaging,* 2022.

[14] J. Chen *et al.*, "3d transunet: Advancing medical image segmentation through vision transformers," *arXiv preprint arXiv:2310.07781,* 2023.

[15] Y. Tang *et al.*, "Self-supervised pre-training of swin transformers for 3d medical image analysis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 20730-20740.

[16] O. Ronneberger, Philipp Fischer, and Thomas Brox., "U-net: Convolutional networks for biomedical image segmentation," *Medical image computing and computer-assisted intervention–MICCAI,* 2015.

[17] R. Azad *et al.*, "Advances in medical image analysis with vision transformers: a comprehensive review," *Medical Image Analysis,* p. 103000, 2023.

[18] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes," *IEEE transactions on medical imaging,* vol. 37, no. 12, pp. 2663-2674, 2018.

[19] X. Yan, H. Tang, S. Sun, H. Ma, D. Kong, and X. Xie, "After-unet: Axial fusion transformer unet for medical image segmentation," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 3971-3981.

[20] L. Chen *et al.*, "Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5659-5667.

[21] L. Li, M. Xu, X. Wang, L. Jiang, and H. Liu, "Attention based glaucoma detection: A large-scale database and CNN model," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 10571-10580.

[22] Y. Fu, J. Liu, and J. Shi, "TSCA-Net: Transformer based spatial-channel attention segmentation network for medical images," *Computers in Biology and Medicine,* vol. 170, p. 107938, 2024.

[23] B. Landman, Z. Xu, J. Igelsias, M. Styner, T. Langerak, and A. Klein, "Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge," in *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, 2015, vol. 5, p. 12.

[24] O. Bernard *et al.*, "Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved?," *IEEE transactions on medical imaging,* vol. 37, no. 11, pp. 2514-2525, 2018.

[25] E. Xie, et al., "SegFormer: Simple and efficient design for semantic segmentation with transformers," *Advances in neural information processing systems,* 2021.

[26] M. Sohaib, S. Shabani, S. A. Mohammed, and B. Parvin, "3D-Organoid-SwinNet: High-content profiling of 3D organoids," *IEEE Journal of Biomedical and Health Informatics,* 2024.

[27] S. Shabani, M. Sohaib, S. A. Mohammed, and B. Parvin, "Multi-Aperture Fusion of Transformer-Convolutional Network (MFTC-Net) for 3D Medical Image Segmentation and Visualization," *arXiv preprint arXiv:2406.17080,* 2024.

[28] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132-7141.

[29] 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.