Letter

Multi-Axis Attention With Convolution Parallel Block for Organoid Segmentation

Pengwei Hu, Xun Deng, Feng Tan, and Lun Hu

Dear Editor,

This letter presents an organoid segmentation model based on multi-axis attention with convolution parallel block. MACPNet adeptly captures dynamic dependencies within bright-field microscopy images, improving global modeling beyond conventional UNet. It excels in sparse global interactions and concurrent computation, yielding enhanced segmentation. MACPNet stands out for its prowess in multi-scale data capture, aligned with diverse distance dependencies inherent in organoid images. Experimental results show that the proposed model outperforms several state-of-the-art methods as well as multiple baseline models in accurate organoid segmentation.

Introduction: Organoids, three-dimensional (3D) cell cultures derived from stem cells or tissue explants, have emerged as a promising tool across various research domains, encompassing developmental biology, disease modeling, and drug screening [1]. These miniaturized and self-organizing tissue structures closely mimic the in vivo organization and function of organs, offering a unique opportunity to study complex biological processes and disease mechanisms in a controlled and reproducible manner [2]. However, unlocking the full potential of organoids necessitates robust and precise methods for their analysis, particularly concerning segmentation and quantification.

Accurate segmentation of organoids from microscopy images is a crucial step in numerous applications, such as tracking cell behaviors, characterizing tissue morphologies, and assessing treatment responses. Traditional manual segmentation is time-consuming, labor-intensive, and prone to variability, limiting its use for large-scale studies. The rapid progress in automated visual analysis provides an opportunity [3]. Therefore, the demand for efficient and reliable automatic segmentation techniques has grown substantially in recent years [4], and organoid structure analysis particularly emphasizes addressing challenges related to its cultivation and dynamic nature.

The cultivation of organoids is a meticulous and time-intensive process that requires skilled professionals to monitor and intervene based on the developmental state of the organoids. This dynamic nature of organoid growth adds an additional layer of complexity to the segmentation process. The evolving morphologies and changing spatial relationships within the organoid structures pose significant challenges for accurate and consistent segmentation. While manual segmentation is considered the gold standard, its high cost [5] makes it impractical for large-scale studies and it lacks the necessary consistency required for rigorous analysis. Therefore, the development of automated and intelligent solutions for organoid segmentation is essential to bridge this gap.

Traditional threshold-based image segmentation methods [6], when applied to organoids, often fall short due to their inability to account for the dynamic changes and intricate internal structures. Machine learning-based approaches have shown promise in various image segmentation tasks but applying them to organoids requires tailored techniques that can adapt to the ever-changing characteristics of the growing tissue. The application of deep learning in the field of medical segmentation [7] has brought inspiration to this task.

Our research addresses the pressing need for automation and intelligence in organoid segmentation by introducing a novel approach called MACPNet that combines multi-axis attention and convolution parallel block. By leveraging these advanced techniques, our method can effectively adapt to the evolving morphology and spatial dynamics of organoids during cultivation. The main contributions of this work are that: 1) MACPNet incorporates an attentional mechanism to capture extended dependencies among pixel points within bright-field microscopy images of organoids. This integration notably boosts the global modeling capability, surpassing the performance of conventional convolution-based UNet organoid segmentation network. 2) MACPNet excels in achieving sparse global interactions in linear time, a notable advancement compared to existing methodologies. This achievement leads to enhanced segmentation performance while concurrently alleviating computational burdens. 3) A distinctive hallmark of MACPNet lies in its concurrent computation of Multi-axis Attention and convolutional techniques. Moreover, it exhibits a remarkable capacity for capturing multi-scale information, replete with diverse distance dependencies intrinsic to organoid images.

Method: Traditional self-attention mechanisms suffer from quadratic complexity, where the computational cost rapidly increases with the size of the input sequence or image, giving rise to issues of excessive attention and computational burden. The fundamental concept behind multi-axis attention is to break down the attention mechanism along spatial axes, resulting in two forms: local and global attention [8].

The block self-attention (Block-SA) uses a chunking strategy to divide the input sequence or image into multiple blocks. Within each block, only local correlation information is considered and correlations with other blocks are ignored, thus reducing computational complexity. Global self-attention (Grid-SA) divides the input sequence or image into a regular grid, then selects a few important elements in the grid and computes the attentional relationships between them and all other elements to form a global distribution of attention. The detailed calculation formula is as follows:

$$
\begin{aligned}
&\text{Block}: (H,W,C) \rightarrow \left(\frac{H}{P}\times P, \frac{W}{P}\times P, C\right) \rightarrow \left(\frac{HW}{P^2}, P^2, C\right)\\
&\text{Grid}: \ (H,W,C) \rightarrow \left(G\times\frac{H}{G}, G\times\frac{W}{G}, C\right)\\
&\qquad \rightarrow \underbrace{\left(G^2, \frac{HW}{G^2}, C\right) \rightarrow \left(\frac{HW}{G^2}, G^2, C\right)}_{\text{swapaxes(axis 1} = -2\text{, axis 2} = -3)}. \qquad (1)
\end{aligned}
$$

As shown in (1), in block attention, the input feature map $x \in \mathbb{R}^{HWC}$ is transformed into a shaped tensor $(H/p \times W/p, p \times p, C)$, representing the partition into non-overlapping windows, with each window's size as $p \times p$. Then, self-attention computation is performed within each window. This operation helps the model focus on local regions, reducing computational complexity while extracting important information within the windows. In grid attention, a fixed $G \times G$ uniform grid is used to gridify the input tensor into dimensions like $G \times G$, $H/G \times W/G$, C. Then, adaptive-sized windows $H/G \times W/G$ are obtained based on these dimensions. Finally, self-attention is calculated on $G \times G$. This sparse global self-attention mechanism has linear complexity, allowing the system to concentrate on a few crucial elements in the global scope and ignore less important elements, thus reducing computational burden.

P. Hu and X. Deng contributed equally to this work. Corresponding author: Lun Hu.

Citation: P. Hu, X. Deng, F. Tan, and L. Hu, "Multi-axis attention with convolution parallel block for organoid segmentation," *IEEE/CAA J. Autom. Sinica*, vol. 11, no. 5, pp. 1295–1297, May 2024.

P. Hu, X. Deng, and L. Hu are with the Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences, Urumqi 830011, the University of Chinese Academy of Sciences, Beijing 100049, and the Xinjiang Laboratory of Minority Speech and Language Information Processing, Urumqi 830011, China (e-mail: hpw@ms.xjb.ac.cn; dengxun22@mails.ucas.ac.cn; hulun@ms.xjb.ac.cn).

F. Tan is with the Merck KGaA, Darmstadt 64293, Germany (e-mail: feng.tan@merckgroup.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JAS.2023.124026

$$\begin{aligned}
&\text{Attention}(Q,K,V) = softmax\left(QK^T/\sqrt{d}\right)V\\
&x \leftarrow \text{x} + \text{Unblock}\,(Attention(\,\text{Block}\,(LN(\,x\,))))\\
&\text{Or } x \leftarrow \text{x} + \text{Ungird}\,(Attention(\,\text{Grid}\,(LN(\,x\,))))\\
&x \leftarrow \text{x} + \text{MLP}\,(\text{LN}(\,x\,))
\end{aligned} \tag{2}$$

where *Q, K, V* are the query, key, and value matrices, and *d* is the hidden dimension [9] in (2). The *Unblock* $(\cdot)$ operation is defined as the reverse operation of the block splitting process described above. *Ungrid* $(\cdot)$ converts the gridded input back to the original two-dimensional feature space. *LN* denotes the layer normalization, where MLP is a standard *MLP* network.

Despite the presence of local attention computation in multi-axis attention, this cannot replace convolution for local feature extraction. We introduce a parallel approach to convolution and attention computation to preserve local feature extraction while computing the global attention mechanism. As shown in Fig. 1, the parallel block of multi-axis attention and convolution serves as the feature extraction backbone module of the network. The parallel features spliced by channels are dimensionally compressed using one-dimensional convolution, and the feature map is output with residual concatenation after extracting similarity information using skip dot product. The computational procedure is shown in (3)

$$\text{out} = \widehat{Z} \otimes \left[\text{Conv1}\,(\text{concat}[\widehat{Z},\widetilde{Z}\,]) \odot \widehat{Z}\right] \tag{3}$$

where $\widetilde{Z}$ represents the output of the convolution and $\widehat{Z}$ represents the feature map calculated by attention; where ⊗ represents the residual connection, and ⊙ is the dot product computation.

The network employs a U-shaped symmetric structure. In the encoder part, the MACP Block facilitates continuous down-sampling of the feature maps until they are 1/32 of the original size. In the decoding part, feature maps with skip connections are combined with the results from the previous decoding layer through channel concatenation and passed through the up-sampling decoding module. Bilinear interpolation is used in the decoding part for 2 × up-sampling, Conv 1 × 1 for channel tuning, and Conv 3 × 3 for feature integration.

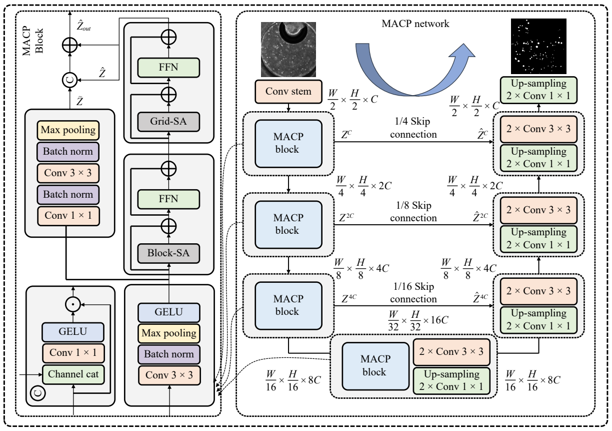

Fig. 1. The overall architecture of MACPNet. MACP Block is a combination of multi-axis attention and convolution parallel block, FFN stands for feedforward neural network, and it includes Block-SA (block self-attention) and Grid-SA (grid self-attention).

$$\widehat{Z}^{\text{iC}} = \text{Dconv3}\,(\text{Up}(\,\text{Conv1}\,(\text{concat}[\,Z^{2\text{iC}},\widehat{Z}^{2\text{iC}}\,]))) \tag{4}$$

where $Z^{2\text{iC}}$ denotes the output of the MACP Block and $\widehat{Z}^{2\text{iC}}$ is the corresponding upsampled output. Dconv3 is the output of a consecutive Conv 3 × 3.

During network training, the batch size for image input is set to 8, with an initial learning rate of 0.01. The learning rate is reduced by half every 50 epochs, and a total of 100 training epochs are performed. Five-fold training is employed, saving the optimal weights on the validation set for each fold. To enhance the network's learning of organoid bright field image features, we employed a multi-loss function training approach, where L_{BCEloss} represents the Binary cross-entropy loss function, and L_{Diceloss} represents the Dice loss function, as shown below:

$$\begin{aligned}
L_{\text{BCEloss}} &= -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \times \log p\,(y_i) + (1-y_i) \times \log\,(1-p\,(y_i))\right]\\
L_{\text{Diceloss}} &= 1 - \left[2\sum_{i=1}^{n}(y_i \times p_i) \Big/ \left(\sum_{i}^{n} y_i + \sum_{i}^{n} p_i\right)\right]
\end{aligned} \tag{5}$$

where $\text{y}_i \in \{0,1\}$ denotes the ground truth pixel values, determining the presence of a specific pixel in the target object, and p_i represents the pixel value predicted by the network, reflecting the network's estimation of whether a pixel belongs to the target object. The final loss function Loss is calculated as follows:

$$\text{Loss} = L_{\text{BCEloss}} + L_{\text{Diceloss}}. \tag{6}$$

Experiments: The organoid training dataset consists of 52 binary images derived from human pancreatic ductal adenocarcinoma (PDAC) organoids [10], with 14 images reserved for validation. To test the model's robustness, three other image types were used: adenoid cystic carcinoma (ACC) of the salivary gland, colon epithelium (Colon), and distal lung epithelia (Lung). In medical imaging tasks, data augmentation effectively enhances the model's generalization ability [11]. This process generated 2000 augmented images by applying random rotations, scaling, elastic distortions, and other morphological transformations. During both network training and inference, the images were converted to single-channel format and resized to a standardized 512 × 512-pixel resolution.

In evaluating the performance of our model, we have selected a set of well-established metrics, namely precision, recall, F1-score, and mean intersection-over-union (mIoU), which are widely utilized in segmentation tasks. Precision quantifies the ratio of true positive predictions to all positive predictions, providing insights into the accuracy of positive predictions. In contrast, recall delineates the proportion of true positive predictions among all instances of ground truth positives, offering a measure of the model's ability to capture relevant instances. The F1-score, a harmonious amalgamation of precision and recall, encapsulates the model's balanced performance. To determine the mIoU, we have taken the average of Intersection over Union (IoU) values across all classes. The IoU is mathematically expressed as TP/(TP + FP + FN). Notably, the mIoU values span the range from 0 to 1, where a value of 1 signifies a perfect alignment between the predicted segmentation and the ground-truth masks. The output results are obtained by averaging the predictions from the best validation weights of each fold in the different models. We compare our method with SOTA segmentation methods. SegNet [12], notable for its employment of a symmetrical encoder-decoder configuration accompanied by pooling indices for up-sampling, stands as one approach. Another approach, A-Unet [13], augments the Unet model with an attentional gate mechanism, which in turn directs the model's focus towards the intended target structures. On a different note, FCN [14] adopts a neural network framework reliant exclusively on convolutional layers, particularly suited for pixel-level semantic segmentation. Conversely, DANet [15] introduces an adaptive integration of local features and global dependencies, a stratagem that lends itself well to segmentation tasks. The scSENet architecture [16], in contrast, introduces concurrent spatial and channel squeeze & excitation methods, concurrently elevating significant features while mitigating the influence of weaker ones. Additionally, OrganoID [10],

Table 1. Network Evaluation on PDAC Segmentation

	PDAC			
Model	mIoU	Precision	Recall	F1
SegNet	0.778	0.740	0.855	0.783
A-Unet	0.801	0.763	0.875	0.806
FCN	0.803	0.769	0.863	0.807
DANet	0.778	0.759	0.813	0.780
scSENet	0.764	0.684	0.875	0.762
OrganoID	0.752	0.702	0.836	0.752
MACPNet	**0.832**	**0.855**	0.862	**0.856**

Table 2. Quantitative Results of Baselines and Our Method on Different Tissues

	ACC				Colon				Lung			
Model	mIoU	Precision	Recall	F1	mIoU	Precision	Recall	F1	mIoU	Precision	Recall	F1
SegNet	0.761	0.742	0.769	0.738	0.664	0.579	0.803	0.630	0.781	0.903	0.730	0.801
A-Unet	0.781	0.645	0.952	0.764	0.791	0.671	0.952	0.783	0.900	0.892	0.946	0.917
FCN	0.795	0.711	0.879	0.780	0.815	0.734	0.911	0.808	0.879	0.907	0.889	0.896
DANet	0.785	0.735	0.810	0.767	0.763	0.717	0.814	0.746	0.831	0.897	0.815	0.852
scSENet	0.797	0.667	0.951	0.782	0.805	0.696	0.942	0.778	0.897	0.869	0.966	0.914
OrganoID	0.766	0.674	0.850	0.745	0.736	0.622	0.866	0.716	0.836	0.794	0.938	0.858
MACPNet	**0.879**	**0.854**	0.928	**0.885**	**0.873**	**0.884**	0.868	**0.872**	**0.926**	**0.918**	**0.974**	**0.945**

tailored for the segmentation of organoids, finds its basis in the Unet architecture. We perform a quantitative evaluation of the prediction results of each network on the PDAC dataset in Table 1. Additionally, we quantitatively evaluate the segmentation results, presented in Table 2. Our evaluation compares the performance of our method against the baselines on different tissues of organoids. The quantitative results show that our approach outperforms the baselines in terms of both robustness and accuracy. And, the segmentation results of our model are shown in Fig. 2.

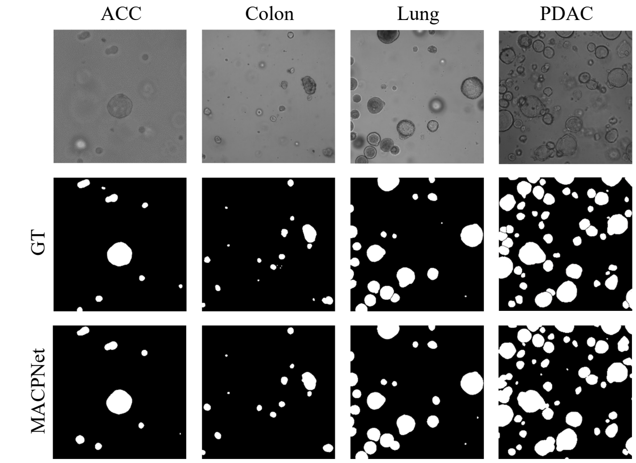

Fig. 2. Visualization comparison of ground truth on the testing dataset.

Conclusions: The expanding frontier of organoid research has ignited remarkable potential across diverse scientific disciplines. These three-dimensional cell cultures faithfully emulate the intricate structures and functions of organs, presenting an unprecedented avenue for probing intricate biological processes and disease pathways. However, unlocking their full potential mandates sophisticated analysis, with a particular emphasis on segmentation and quantification. Although manual methodologies offer accuracy, they falter when scalability and consistency are demanded, thereby obstructing expansive studies. To surmount this obstacle, we introduce MACPNet, a pioneering framework that amalgamates multi-axis attention and convolution parallel block. By synergizing these cutting-edge techniques, our approach seamlessly adapts to the ever-shifting morphology and spatial dynamics intrinsic to organoid growth. Notable contributions encompass the amplified global modeling prowess via attentional mechanisms, the achievement of linear time sparse global interactions, and the comprehensive capture of multi-scale information. This effectively surpasses the limitations of conventional segmentation methods. In essence, our efforts drive the frontiers of automation and intelligence in organoid analysis, thereby enabling deeper insights and broader applications across developmental biology, disease modeling, and drug screening.

Acknowledgments: This work was supported by the Xinjiang Tianchi Talents Program (E33B9401), the Natural Science Foundation of Xinjiang Uygur Autonomous Region (2023D01E15), the National Natural Science Foundation of China (62302495), and the National Natural Science Foundation of China (62373348).

References

[1] K. Kretzschmar and H. Clevers, "Organoids: Modeling development and the stem cell niche in a dish," *Dev. CELL*, vol. 38, no. 6, pp. 590–600, Sept. 2016.

[2] A. Fatehullah, S. H. Tan, and N. Barker, "Organoids as an in vitro model of human development and disease," *Nat. CELL Biol.*, vol. 18, no. 3, pp. 246–254, Mar. 2016.

[3] T. Wang, X. Xu, F. M. Shen, and Y. Yang", "A Cognitive memory-augmented network for visual anomaly detection," *IEEE/CAA J. Autom. Sinica*, vol. 8, no. 7, pp. 1296–1307, Jul. 2021.

[4] C. Lee, H. Hasegawa, and S. Gao, "Complex-valued neural networks: A comprehensive survey," *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 8, pp. 1406–1426, Aug. 2022.

[5] P. Huang, J. Han, N. Liu, J. Ren, and D. Zhang, "Scribble-supervised video object segmentation," *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 2, pp. 339–353, Feb. 2022.

[6] S. Pare, A. Kumar, V. Bajaj, and G. K. Singh, "A context sensitive multilevel thresholding using swarm based algorithms," *IEEE/CAA J. Autom. Sinica*, vol. 6, no. 6, pp. 1471–1486, Nov. 2019.

[7] Y. Xia, H. Yu, and F.-Y. Wang, "Accurate and robust eye center localization via fully convolutional networks," *IEEE/CAA J. Autom. Sinica*, vol. 6, no. 5, pp. 1127–1138, Sept. 2019.

[8] Z. Tu, H. Talebi, H. Zhang, *et al*., "MaxViT: Multi-axis vision transformer," in *Computer Vision – ECCV 2022*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds., Cham, Switzerland: Springer Nature Switzerland, 2022, pp. 459–479.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Advances in Neural Information Processing Systems 30: Annu. Conf. Neural Information Processing Systems*, Long Beach, USA, 2017, pp. 6000–6010.

[10] J. M. Matthews, B. Schuster, S. S. Kashaf, *et al*., "OrganoID: A versatile deep learning platform for tracking and analysis of single-organoid dynamics," *PLOS Comput. Biol.*, vol. 18, no. 11, p. e1010584, 2022.

[11] A. K. Bhandari, A. Ghosh, and I. V. Kumar, "A local contrast fusion based 3D OTSU algorithm for multilevel image segmentation," *IEEE/CAA J. Autom. Sinica*, vol. 7, no. 1, pp. 200–213, Jan. 2020.

[12] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, 2017.

[13] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, and Rueckert. "Attention U-NET: Learning where to look for the pancreas," arXiv preprint arXiv: 1804.03999, 2018.

[14] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Boston, USA, 2015, pp. 3431–3440.

[15] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2019, pp. 3146–3154.

[16] A. G. Roy, N. Navab, and C. Wachinger, "Concurrent spatial and channel 'squeeze & excitation' in fully convolutional networks," in *Proc. Medical Image Computing and Computer Assisted Intervention*, 2018.