AC-MLP: AXIAL CONVOLUTION-MLP MIXER FOR NUCLEI SEGMENTATION IN HISTOPATHOLOGICAL IMAGES

Nguyen Thanh Thu, Dinh Binh Duong, Tran Thi Thao*, Pham Van Truong

School of Electrical and Electronic Engineering, Hanoi University of Science and Technology, Vietnam

*Corresponding author: thao.tranthi@hust.edu.vn

(Received: July 11, 2024; Revised: September 01, 2024; Accepted: September 27, 2024)

DOI: 10.31130/ud-jst.2024.332E

Abstract - Recent MLP-Mixer has a good ability to handle long-range dependencies, however, to have a good performance, one requires huge data and expensive infrastructures for the pre-training process. In this study, we proposed a novel model for nuclei image segmentation namely Axial Convolutional-MLP Mixer, by replacing the token mixer of MLP-Mixer with a new operator, Axial Convolutional Token Mix. Specifically, in the Axial Convolutional Token Mix, we inherited the idea of axial depthwise convolution to create a flexible receptive field. We also proposed a Long-range Attention module that uses dilated convolution to extend the convolutional kernel size, thereby addressing the issue of long-range dependencies. Experiments demonstrate that our model can achieve high results on small medical datasets, with Dice scores of 90.20% on the GlaS dataset, 80.43% on the MoNuSeg dataset, and without pre-training. The code will be available at https://github.com/thanhthu152/AC-MLP.

Key words - Depthwise convolution; MLP-Mixer; Nuclei segmentation; Token mixing

1. Introduction

The distribution and density of nuclei in the tissues are important and necessary markers in cancer diagnosis. Therefore, the detection and segmentation of cell nuclei is getting more and more attention and is an essential task in biomedical engineering. Nuclei segmentation aids in tissue structure determination, cell growth analysis, and the study of cell responses to environmental changes. On this basis, researchers can derive cell characteristics, diagnose disease severity, and research drugs.

Studies on cell nuclei identification have been around for a long time. First of all, there is the appearance of the microscope, which allows people to see cells with a microscopic size. Microscopy supports the acquisition of images of cells, from which traditional methods such as threshold-based, region-based, and edge-based are widely utilized to segment cell nuclei. However, the nuclei often have tiny sizes and high distribution densities, which causes many difficulties for the traditional segmentation techniques.

The emergence of deep learning techniques in the field of Computer Vision brings a novel approach to image segmentation, particularly in the realm of medical imaging. The presence of Convolutional Neural Networks (CNNs) is a leap and opens a series of related studies in image processing. In 2015, the U-Net model was introduced, which utilized a U-shaped architecture to effectively segment medical images. This model extracts multi-scale context information through its encoder and reconstructs the input size through its decoder, while skip connections are used to avoid information loss. Since then, many variants of this architecture have been proposed to improve performance. Some typical models can be mentioned as Double Unet [2], Attention Unet [3], Unet++ [4], ResUnet ++ [5], etc. On the other hand, concerned with computational cost and real-life applications, researchers began to focus on lightweight models. Therefore, depthwise separable convolutions are now more commonly used as alternatives for conventional convolutions to reduce the number of parameters while still achieving high performance. For example, DSCA-Net [6] combines depthwise separable convolution with an attention mechanism in a U-shape architecture to create a lightweight network for accurate medical image segmentation. MobileNets [7] is another lightweight model which employed depthwise convolutions and was successfully embedded in mobile visual applications. Most recently, U-Lite [8] was introduced as an effective model with less than 1 million parameters, using axial depthwise convolution, which can give promising results on medical datasets.

Recently, the arrival of Vision Transformer [9] has attracted a lot of research and overwhelmed CNNs dominance in computer vision. Vision Transformer (ViT) applied the Transformer from the Natural Language Processing (NLP) domain to Computer Vision by dividing an image into many patches and flattening them to vectors before taking them as input of the Transformer. With the self-attention mechanism, ViT has a better ability to learn long-range dependencies and extract global context information than CNNs. However, Transformer has a large amount of computation and may require a lot of resources during the training process. Swin Transformer [10] was proposed then to reduce the computational cost by limiting self-attention computation to non-overlapping local windows. Besides Transformers, Multi-layer Perceptrons (MLPs) have also been applied to vision tasks, but are less common. MLP-Mixer inherits the patch partition of ViT and passes embedded patches through several layers of the token mixer and channel mixer. Both these mixing operators utilize pure MLP, while they can still achieve comparable results to those of CNN-based models and transformer-based models on classification tasks. Similar to ViT, MLP-Mixer is also successful in extracting global information from input images. However, both require a large enough dataset to achieve good performance.

In the field of medical image segmentation, obtaining a large data set, particularly for nuclei, can be challenging. In the nuclei segmentation task, because the density of cells is very thick and the size of a cell is small, experts must

take great care and spend a significant amount of time accurately segmenting cell nuclei in order to produce a high-quality dataset. Moreover, because of the small size of nuclei, local receptive fields may help aggregate context information from nuclei better. To address these challenges, we develop a new model that leverages ideas from the MLP-Mixer and axial depthwise separable convolution of the U-Lite model. By using axial depthwise separable convolution to replace the MLP-Mixer token mix, our proposed model has achieved the following contributions:

- Propose Axial Convolution Mixer module based on the concept of MLP-Mixer.
- Propose AC-MLP model for image segmentation task with two branches of the encoder and adapt attention to the decoder.
- Experimentally, AC-MLP achieves the state-of-the-art (SOTA) performance on small datasets like GlaS and MoNuSeg using a low number of parameters without pre-trained.

2. Related work

2.1. Axial Depthwise Convolution

U-Lite [8] takes the usage of axial depth-wise convolution as the main operator to aggregate spatial information of the feature maps. This module is established by simply taking the sum of two separated operators 1×7 and 7×1 depth-wise convolution, axial depth-wise convolution does not overly increase the number of learnable parameters as well as the model's complexity. However, it can result in a comparable or even slightly better performance due to the inductive bias in the receptive field.

2.2. MLP-Mixer

Besides Transformers and CNNs, MLP-like models have been widely used and are known as a recently emerging paradigm for Computer Vision. MLP-Mixer is the first study that utilized pure MLP as token-mixers on spatial and channel representations of the feature map. Specifically, for the image classification task, the image is first passed through a per-patch fully-connected layer for patch embedding. After that, they are adapted to several numbers of mixer layers. Each layer comprises two stages separately, one is Token-mixer for spatial feature extraction, and the remaining one is responsible for Channel-mixing, i.e., encoding features along their channel dimension. Finally, a classification header is designed in the last layer for the classification tasks. Despite having a very straightforward architecture, MLP-Mixer can achieve promisingly comparable results on ImageNet's classification benchmarks. This new paradigm has inspired various architectures for performance improvements, including ViP [12], CycleMLP [13], and AS-MLP [14].

2.3. Progressive Atrous Pyramid Pooling (PASPP)

PASPP [15] utilizes multiple atrous convolutional layers with different dilation rates and progressive concatenations to capture multi-scale representations of an object in feature maps. It has been experimented that this module can impressively achieve a better performance on image segmentation tasks compared to the previously proposed module ASPP [16]. Based on the belief that the larger the dilation rate is, the more global information the model can aggregate, PASPP not only provides a larger receptive field to the model compared to traditional 3×3 convolutions but also retains the number of computational parameters within a limited resource.

2.4. Convolutional Block Attention Module (CBAM)

Inspired by Squeeze and Excitation [17], CBAM [18] is regarded as a lightweight efficient attention module that considerably improves the performance of CNN architectures. A CBAM contains two main stages: channel attention and spatial attention. The channel attention module applies max-pooling and average pooling on every spatial dimension to aggregate important information per each channel of the feature maps. They are next passed through double fully-connected layers, a sigmoid layer, and then multiplied again with the input. This mechanism helps to determine which channels are more important than other ones. The spatial attention module is implemented similarly, however, 7×7 convolutions are utilized instead of fully-connections due to the limitation of calculation, which allows the model to precisely focus on spatial representations of the feature maps.

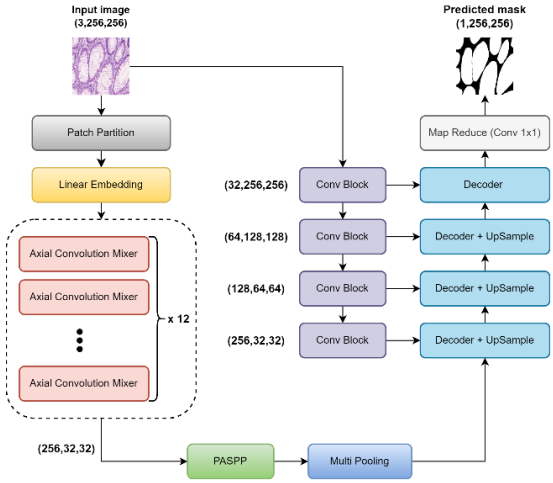
3. Methodology

Inspired by the architecture of Axial Attention MLP-Mixer [19], in this study, we proposed a similar architecture with some new enhancements, namely Axial Convolution-MLP Mixer (AC-MLP). The specific architecture of our proposed model is shown in *Figure 1*. Given an input image $I_o \in \mathbb{R}^{3\times H\times W}$. The encoder of our proposed model consists of two branches to extract features. In the first branch, the image is divided into many non-overlapping patches of size $p \times p$. After that, all patches are projected to vectors of the same size and become the input of a network with 12 successive Axial Convolution Mixer blocks to produce context information. The output of this network undergoes a bottleneck block comprising a PASPP module and Multi-Pooling layers, working together to synthesize beneficial features. Parallel to the first branch, in the second branch, the input image is passed through consecutive Conv Block (*Figure 4b*) and MaxPooling layers to extract context information on multiple scales. Finally, after having the features from the input image, we combine these features in both branches of the model and pass them through the decoder and upscale to select information and reconstruct the original size of the input image before giving the final prediction mask.

3.1. Proposed Axial Convolution Mixer

The effectiveness of Axial Depthwise Convolution throughout the U-Lite model has motivated us to exploit this module to develop a novel spatial mixing operator, serving as a replacement for the token mixing mechanism of the MLP-Mixer, with the aim of adapting to the task of nucleus segmentation. While global information is undoubtedly important in image segmentation, we hypothesize that, for nucleus segmentation, local

information and dense granularity are the features that warrant greater focus. Consequently, the parallel use of horizontal and vertical kernels, as implemented in Axial.



Firgure 1. *General structure of the proposed model*

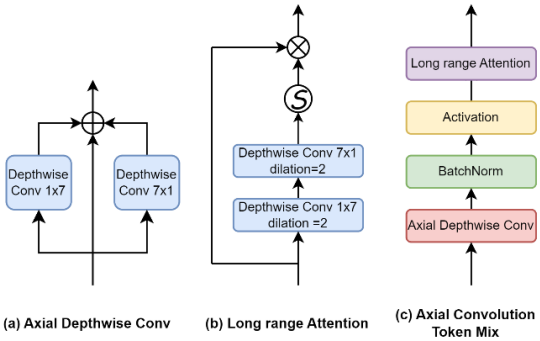


Figure 2. *Proposed Axial Convolution Token Mixing module*

Depthwise Convolution, offers a more flexible approach to learning local information, in contrast to utilizing MLP networks to capture global spatial relationships between pixels, as in the token-mixer architecture of the MLP-Mixer. Furthermore, compared to Transformer-based models, Axial Depthwise Convolution enhances the focus on learning local information along the horizontal and vertical dimensions of pixels, which aligns well with the compact and densely structured nature of nuclei. To address the limitations in learning long-range dependencies, we propose a long-range attention module (Figure 2b) subsequent to Axial Depthwise Convolution, employing dilated convolution to expand the receptive field. The mathematical formulation of this module is presented as follows:

$$y = \text{GELU}(\text{BN}(\text{ADC}(x))) \quad (1)$$

$$z = y \times \text{Sigmoid}\left(\text{DW}_{1\times7,r=1}\left(\text{DW}_{7\times1,r=2}(y)\right)\right) \quad (2)$$

where, x is the input features with the shape $C \times \frac{H}{p} \times \frac{W}{p}$; z is the output features; BN, ADC and DW stand for Batch Normalization, Axial Depthwise Convolution and Depthwise Convolution, respectively, and r is the dilation rate of the convolution. The new spatial mixer mentioned above is used to replace token mixer of MLP-Mixer architecture and we have a new module, namely Axial Convolution Mixer as show in Figure 2 and Figure 4a.

3.2. Bottleneck Block

In the bottleneck of AC-MLP, we utilize the PASPP [15] and Multi-Pooling to capture multi-scale representations of high-level feature maps. As depicted in *Figure 3*, the Multi-Pooling uses three max-pooling layers with different sizes of kernel $k = 2, 4$ and 8. These max-pooled features are passed through point-wise convolutions and then interpolated back to the original shape. Finally, we concatenate them with the output of PASPP, and one gain feed the encoded features to 1×1 convolution to recover the features' dimension. Different from PASPP, Multi-Pooling can focus on the most representative characteristics of an image, surpass non-essential information, and thus give the model an inductive bias. The intuition here is that depending on each image and data, the nuclei may have different shapes and they can stand alone or gather in groups, then a combination of convolution and max-pooling operators, where the kernel size varies flexibly, is an effective way to detect them.

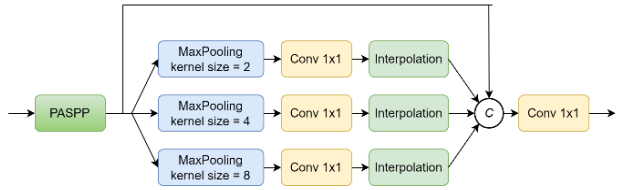


Figure 3. *Bottleneck design of AC-MLP model*

3.3. Decoder Block

After the initial feature extraction process, wherein feature maps are obtained from two branches of the encoder, we propose a novel decoder architecture that capitalizes on an attention mechanism, as illustrated in Figure 4c. To begin, we concatenate the feature maps derived from the two branches. Subsequently, a pointwise convolution operation is employed to transform and adjust the channel dimensions, effectively enhancing the subsequent processing. Inspired by the Channel and Spatial Attention Module (CBAM) concept, our approach differentiates itself by reimagining the arrangement of Channel Attention and Spatial Attention into parallel blocks, in contrast to the linear configuration seen in CBAM. This innovative parallel configuration help the model to extract and emphasize crucial insights from both the spatial and channel dimensions of the feature maps. The outputs of these parallel blocks are then concatenated, facilitating the cohesive integration of the identified salient features. This unified representation then undergoes convolution, utilizing a 3×3 kernel, as a pivotal step in the subsequent stages of feature refinement.

4. Experiment

4.1. Implementation Detail

4.1.1. Dataset

To evaluate the effectiveness and efficiency of our proposed method, we utilize two histopathological nuclei datasets for the image segmentation task. The Gland Segmentation (GlaS) dataset [20] introduced in the Colon Histology Images Challenge Contest, was created to promote research in segmentation algorithms on images of

hematoxylin and Eosin (H&E) stained slides. This dataset consists of 74 benign images and 91 malignant images in total, which is divided into 85 images for training and 80 images for testing. Multi-organ Nucleus Segmentation (MoNuSeg) dataset [21], another nuclei dataset, aims to look for the best nuclei segmentation techniques on a diverse set of H&E stained histology images obtained from multiple organs of patients. This dataset includes 30 images for training and 14 images for testing, with nearly 30,000 nuclear boundary annotations. In our experiment, all the images are resized to 256×256. Furthermore, the training images are pre-processed before feeding to the model through augmentation techniques including image rotating, horizontal flipping, and vertical flipping to enrich the training data and avoid over-fitting.

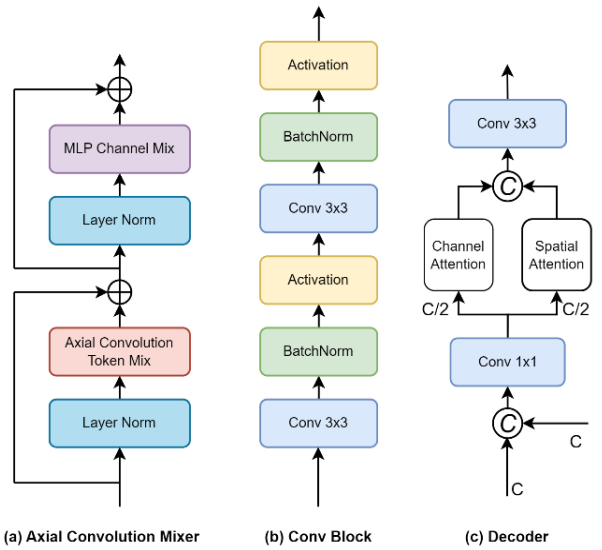


Figure 4. *The structure of Axial Convolution Mixer, Double Convolution and Decoder Block*

4.1.2. Training strategy

We conducted the proposed model on the PyTorch framework running on NVIDIA Tesla T4 GPU with 16GB of memory. The training process was implemented on 100 epochs with a batch size of 16, where the learning rate was initialized at 0.001 and decayed by a factor of 2 after every 10 epochs. During the training phase, we adopted the Adam optimizer [22] with the composite loss function between Binary Cross-Entropy (BCE) loss and Dice loss. Define $1 \times H \times W$ as the shape of the predicted mask and $N = H \times W$ presents its total number of pixels. The loss function employed in the experiment is presented as follows:

$$L_{\text{Total}}(y, \hat{y}) = \gamma L_{\text{BCE}}(y, \hat{y}) + (1 - \gamma) L_{\text{Dice}}(y, \hat{y}) \quad (3)$$

$$L_{\text{BCE}}(y, \hat{y}) = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)] \quad (4)$$

$$L_{\text{Dice}} = 1 - \frac{2 \sum_{i=1}^{N} y_i \hat{y}_i}{\sum_{i=1}^{N} (y_i + \hat{y}_i) + \varepsilon} \quad (5)$$

Where, $y = \{y_1, y_2, \ldots, y_N\}$ and $\hat{y} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_N\}$ are respectively sets of pixels on the ground truth and the predicted mask, where $y_i \in \{0, 1\}$ and $\hat{y}_i \in (0, 1)$ for all $i = 1, 2, \ldots, N$. Besides, ε was added to avoid the zero-denominator. In our experiment, $\gamma = 0.5$, and $\varepsilon = 1.0$.

4.1.3. Evaluation metric

To quantitatively evaluate the performance, we utilized Dice Similarity Coefficient (Dice) and Intersect over Union (IoU) metrics, which are standard evaluation indicators typically used for calculating the overlap between the ground truth and the predicted mask. The mathematical representations of Dice and IoU are expressed as follows:

$$\text{Dice} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \quad (6)$$

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \quad (7)$$

where TP, FP, FN respectively stand for True Positives, False Positives and False Negatives between the ground truth and the prediction of an image.

4.2. Representative Results

Figure 5 and Figure 6 show some segmentation results of AC-MLP model on GlaS and MoNuSeg datasets, respectively. It is observed that the predictions from our model match well with the ground truths. Specifically, the model can properly detect the boundary of each nucleus on GlaS, and maintain the spatial arrangement of the multi-organ nuclei on MoNuSeg, even though they are more numerous and varied.

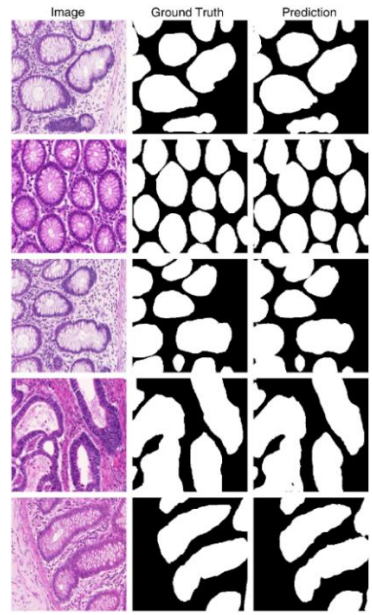


Figure 5. *Some representative results of AC-MLP for Gland Segmentation (GlaS)*

4.3. Comparative Results

Table 1 and Table 2 evaluate the quantitative results on two nuclei datasets GlaS and MoNuSeg, respectively. We compared our model with other state-of-the-art architectures including both the CNN-based and Transformer-based models. As can be seen from the tables, our proposed method outperforms the previously proposed models in terms of Dice and IoU metrics on both datasets. Specifically, AC-MLP reaches Dice scores of 90.20% on the GlaS dataset and 80.43% on the MoNuSeg dataset. This demonstrates the effectiveness of our approach in accurately segmenting nuclei in histopathology images.

In Table 3, we further demonstrate the efficiency of the AC-MLP model, where its Dice and IoU results surpass those of the other MLPs-based variants. One noticeable realization is that AC-MLP has significantly fewer parameters than MLP-Mixer, with a total of 8.5M parameters, while still delivering good performance.

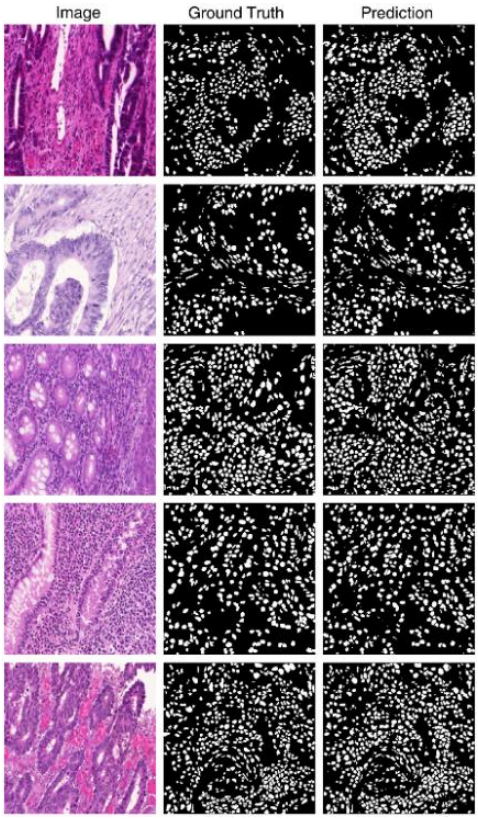


Figure 6. *Some representative results of AC-MLP for Multi-organ Nucleus Segmentation (MoNuSeg)*

Table 1. *Quantitative evaluation results on GlaS dataset compared to previously proposed model*

Type	Model	Dice	IoU
CNNs baselines	UNet [1]	77.78	65.34
	Unet++ [4]	78.03	65.55
	ConvUNeXt [23]	78.04	64.42
	UneXt [24]	86.49	77.77
Transformers baselines	Axial Attn Unet [19]	76.30	63.03
	MedT [25]	81.02	69.61
	Swin Unet [26]	88.25	79.86
Ours	**AC-MLP**	**90.20**	**82.89**

Table 2. *Quantitative evaluation results on MoNuSeg dataset compared to previously proposed models*

Type	Model	Dice	IoU
CNNs baselines	UNet [1]	76.45	62.86
	Unet++ [4]	77.57	66.20
	ConvUNeXt [23]	73.70	60.07
	UneXt [24]	78.04	64.42
Transformers baselines	Axial Attn Unet [19]	76.83	62.49
	MedT [25]	79.55	64.42
	Swin Unet [26]	78.49	64.72
Ours	**AC-MLP**	**80.43**	**67.46**

Table 3. *Quantitative comparisions with variants of MLPs on GlaS dataset*

Methods	Patch size	Depth (layer)	Params (M)	Dice	IoU
MLP-Mixer [11]	16	24	100	82.83	70.81
Permutator [12]	8	36	74	84.21	72.80
AxialAtt-MLP [19]	8	24	29	84.99	73.97
Ours	**8**	**12**	**8.5**	**90.20**	**82.89**

5. Conclusion

In this paper, we leverage the primary knowledge about Axial Depthwise Convolution and MLP-Mixer architecture to propose a new model AC-MLP for histopathological nuclei image segmentation. Our novel module, Axial Convolution Token Mixing, is designed to capture large-scale information and preserve long-range dependencies. Experimental results show that AC-MLP can achieve SOTAs performance meanwhile having an efficient number of computational parameters. In the future, we will thoroughly consider the encoder's CNN-based branch to further improve the model's performance for the large use case of medical images on the segmentation task.

Acknowledgement: This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 102.05-2021.34.

REFERENCES

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation", *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, Proceedings, Part III 18. Springer International Publishing*, 2015.

[2] D. Jha, M. A. Riegler, D. Johansen, P. Halvorsen, and H.D. Johansen, "Doubleu-net: A deep convolutional neural network for medical image segmentation", *2020 IEEE 33rd International symposium on computer-based medical systems (CBMS)*, 2020.

[3] O. Oktay *et al.*, "*Attention u-net: Learning where to look for the pancreas*", arXiv preprint arXiv:1804.03999, 2018.

[4] Z. Zhou, M.M.R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation", *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4. Springer International Publishing*, 2018.

[5] D. Jha *et al.*, "Resunet++: An advanced architecture for medical image segmentation", 2019 IEEE International Symposium on Multimedia (ISM). IEEE, 2019.

[6] T. Shan, J. Yan, X. Cui, and L. Xie, "DSCA-Net: A depthwise separable convolutional neural network with attention mechanism for medical image segmentation", Math Biosci Eng., vol. 20, pp. 365-382, 2022.

[7] AG. Howard *et al.*, "*Mobilenets: Efficient convolutional neural networks for mobile vision applications*", arXiv preprint arXiv:1704.04861, 2017.

[8] D.B. Dinh, T.T Nguyen, T.T. Tran, and V.T Pham, "1M parameters are enough? A lightweight CNN-based model for medical image segmentation*", 2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE*, 2023.

[9] A. Dosovitskiy *et al.*, *"An image is worth 16x16 words: Transformers for image recognition at scale"*, arXiv preprint arXiv:2010.11929, 2020.

[10] Z. Liu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows", *Proceedings of the IEEE/CVF international conference on computer vision*, 2021.

[11] O. Tolstikhin *et al.* "*Mlp-mixer: An all-mlp architecture for vision*", Advances in neural information processing systems 34, 2021, pp.24261-24272

[12] Q. Hou, Z. Jiang, L. Yuan, M.M. Cheng, S. Yan, and J. Feng, "Vision permutator: A permutable mlp-like architecture for visual recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*. vol.45, no.1, pp. 1328-1334, 2022.

[13] S. Chen, E. Xie, C. Ge, R. Chen, D. Liang, and P. Luo, "*Cyclemlp: A mlp-like architecture for dense prediction",* arXiv preprint arXiv:2107.10224, 2017.

[14] D. Lian, Z. Yu, X. Sun, and S. Gao, "*As-mlp: An axial shifted mlp architecture for vision*", arXiv preprint arXiv:2107.08391, 2021.

[15] Q. Yan *et al., "COVID-19 chest CT image segmentation--a deep convolutional neural network solution",* arXiv preprint arXiv:2004.10987, 2020.

[16] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs", *IEEE transactions on pattern analysis and machine intelligence*. Vol. 40, no. 4, pp. 834-848, 2017.

[17] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks", *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018.

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

[19] H. P. Lai, T.T Tran, and V.T Pham, "Axial attention mlp-mixer: A new architecture for image segmentation", *2022 IEEE Ninth International Conference on Communications and Electronics (ICCE). IEEE*, 2022.

[20] K. Sirinukunwattanan *et al.* "Gland segmentation in colon histology images: The glas challenge contest", *Medical image analysis*, vol. 35, pp. 489-502, 2017.

[21] N. Kumar *et al*., "A multi-organ nucleus segmentation challenge", *IEEE transactions on medical imaging*, vol. 39, no. 5, pp. 1380-1391, 2019.

[22] K. Kingma and J. Ba, *"Adam: A method for stochastic optimization",* ArXiv Preprint ArXiv:1412.6980, 2014.

[23] Z. Han, M. Jian, and G.G. Wang, "ConvUNeXt: An efficient convolution neural network for medical image segmentation". *Knowledge-Based Systems*, vol. 253, no. C, 2022.

[24] J. J. M. Valanarasu, and V. M. Patel, "Unext: Mlp-based rapid medical image segmentation network", *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22*, 2022.

[25] J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. Patel, "Medical transformer: Gated axial-attention for medical image segmentation". *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2021. pp. 36-46.

[26] H. Cao *et al., "Swin-unet: Unet-like pure transformer for medical image segmentation", European conference on computer vision. Cham: Springer Nature Switzerland*, 2022.

[27] Z. Liu, H. Mao, C. Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s", *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022.