2022 7th International Workshop on Big Data and Information Security (IWBIS) | 978-1-6654-8950-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/IWBIS56557.2022.9924685

Modified MultiResUNet for Left Ventricle Segmentation from Echocardiographic Images

Fityan Azizi*, Akbar Fathur Sani*, Rinto Priambodo*, Wisma Chaerul Karunianto*,
Mgs M Luthfi Ramadhan*, Muhammad Febrian Rachmadi*†, and Wisnu Jatmiko*
*Faculty of Computer Science, Universitas Indonesia, Depok, Indonesia
†RIKEN research, Tokyo, Japan
Email: fityan.azizi, akbar.fathur, rinto.priambodo11, wisma.chaerul11, mgs.m01{@ui.ac.id},
febrian.rachmadi@riken.jp, wisnuj@cs.ui.ac.id

***Abstract*—An accurate assessment of heart function is crucial in diagnosing the cardiovascular disease. One way to evaluate or detect the disease can use echocardiography, by detecting systolic and diastolic volumes. However, manual human assessments can be time-consuming and error-prone due to the low resolution of the image. One way to detect heart failure on echocardiogram is by segmenting the left ventricle on the echocardiogram using deep learning. In this study, we modified the MultiResUNet model for left ventricle segmentation in echocardiography images by adding Atrous Spatial Pyramid Pooling block and Attention block. The use of multires blocks from MultiResUnet is able to overcome the problem of multi-resolution segmentation objects, where the segmentation objects have different sizes. This problem has similar characteristics to echocardiographic images, where the systole and diastole segmentation objects have different sizes from each other. Performance measure were evaluated using Echonet-Dynamic dataset. The proposed model achieves dice coefficient of 92%, giving an additional 2% performance result compared to the MultiResUNet.**

***Index Terms*—Heart Function, Echocardiography, Semantic Segmentation, Deep Learning**

I. Introduction

Cardiovascular disease is a dangerous disease with the highest mortality rate in the world [1]. In detecting the disease, an accurate assessment of heart function is crucial in diagnosing the disease. One way to evaluate or detect the disease is to use echocardiography, i.e. by detecting systolic and diastolic volumes [2]. However, manual human assessments can be time-consuming and error-prone due to the low resolution of the image. Human assessment of cardiac function focuses on taking a limited sample of the cardiac cycle and has considerable inter-observer variability. It is very important for automatic detection to be carried out so that checking is carried out effectively and reduces errors.

One way to detecting heart failure on echocardiogram is by segmenting the left ventricle on the echocardiogram using deep learning [2]. U-net is a neural network architecture designed primarily for image segmentation. The basic structure of a U-net architecture consists of two paths. The first path is the encoder path, which is similar to a regular convolution neural network (CNN) and provides classification information. The second is the decoder path, consisting of up-convolutions and concatenations with features from the encoder path. This expansion allows the network to learn localized classification information. In addition, the expansion path also increases the resolution of the output, which can then pass to a final convolutional layer to create a fully segmented image. U-Net architecture and its development has been widely used to segment biomedical images such as liver, skin, and blood vessels. A number of studies show that this model has a good performance in segmentation task [4].

To support research related to cardiac function assessment, an echocardiographic image dataset is available called Echonet-Dynamic, which consists of 10036 video echocardiograms and annotations from experts. Figure 1 is an example of an image in Echonet-Dynamic, which consists of a series of videos visualizing the heart from different angles and positions. This dataset has been used to segment the left ventricle using the DeeplabV3-Resnet50 model [3].

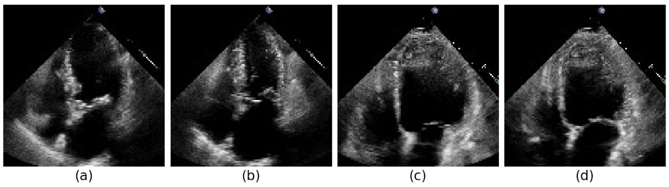

Fig. 1. Echocardiographic images in Echonet-Dynamic Dataset [2]

One way to assess cardiac function is to measure the Left Ventricle Ejection Fraction (LVEF). LVEF is obtained from the calculation between End Systolic Volume (ESV) and End Diastolic Volume (EDV), which is calculated by (EDV - ESV) / EDV and expressed as a ratio. Cardiac function is considered healthy if the LVEF value is more than 50 [2]. Figures 1(a) and 1(b) are examples of healthy cardiac function with an LVEF value of 78, where there is a significant volume change between EDV at 1(a) and ESV at 1(b). In contrast to 1(a) and 1(b), figures 1(c) and 1(d) are examples of unhealthy cardiac function with an LVEF value of 24. There was no large volume difference between the EDV at 1(c) and the ESV at 1(d).

In segmenting medical images using U-Net-based architecture, Ibtehaz and Rahman [5] made modifications to the U-Net architecture by replacing the main block in U-Net into a

978-1-6654-8950-8/22/$31.00 ©2022 IEEE

multi-residual block. With multi-residual block, the model is able to overcome the problem of multi-resolution objects or segmentation objects with different scales or sizes. The model, named MultiResUNet, also replaces the skip connection in U-Net with a residual path to overcome the semantic gap problem between the encoder and the decoder. The model produces a significant jaccard index value compared to the U-Net architecture, by testing five different types of medical datasets. Jaccard index is defined as the ratio of the intersection and union of the two sets [5].

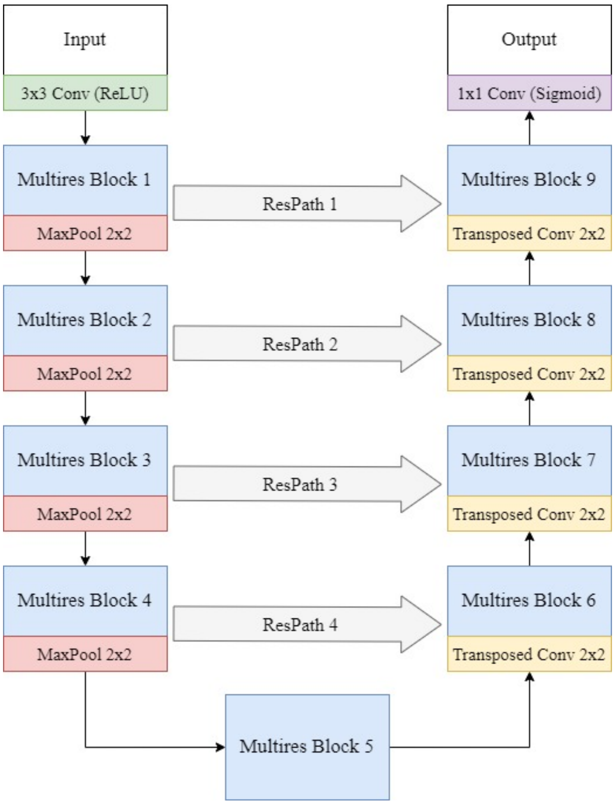

Fig. 2. The architecture of the MultiResUNet Model (Reproduced).

Jha et al. [6] modified the U-Net architecture by replacing the main U-Net block into a residual block added with Squeeze and Excitation Net. In the decoder, an attention block is added to each residual block which is claimed to be able to increase the performance of the model. The study also added the Atrous Spatial Pyramid Pooling block as a bridge between the encoder and decoder. The addition of these blocks is considered capable of filtering information at different scales. The study resulted in a significant performance increase compared to previous studies, by obtaining a dice coefficient value of 0.81 in the Polyp dataset.

Amer et al. [7] modify the U-Net model by making changes to the main block which is changed to the residual block and Squeeze and Excitation Net. The study also added a dilated convolution block which is claimed to be able to overcome the problem of different left ventricular segmentation objects. The model is used to segment the left ventricle of the heart on echocardiography images on the CAMUS dataset. The study achieved a dice coefficient value of 0.95 and provides improved performance compared to the U-Net architecture.

In this study, we modified the MultiResUNet model for left ventricle segmentation in echocardiography images by adding features from the previously described related studies. The addition of features is based on research conducted by [6] and [7] in using dilated convolution to filter features on different segmentation objects and using attention blocks to improve model performance. We chose MultiResUNet as the main architecture in this study because the model uses multires block that is made to overcome the problem of segmenting objects that have different resolutions. This has similarities with echocardiography images, where the size of the left ventricle object varies. The main contributions of this study are summarized as follows:

- Add the Attention Block in the decoder and the ASPP Block as part of the bridge between encoder and decoder.
- Change the activation function to SELU activation function.
- Compare the modified model to original model and several different models on Echonet-dynamic Dataset.

II. Proposed Model

The main architecture in this research is the MultiResUNet architecture. The use of multires blocks is able to overcome the problem of multi-resolution segmentation objects, where the segmentation objects have scale variations or different sizes. This problem also has similar characteristics to the systole and diastole object in echocardiographic images, where segmentation objects have different sizes from each other. In general, the model architecture proposed in this study is to add ASPP blocks and Attention blocks to the MultiResUNet model architecture. Figure 3 shows the architecture in this study, the changes made compared to the MultiResUNet architecture in Figure 2 are the addition of an ASPP block to the bridge and an attention block before upsampling on each multires block in the decoder.

A. *Multires Block*

CNN has a limited tolerance for scale variations [8]. This will affect the segmentation results. One way to deal with scale variations on CNN in U-Net architecture is to replace the convolutional layers with inception like blocks from inception net [5]. Multires block, that is built to solve scale variations problems in object segmentation is inspired by inception net, where the block uses three different convolution kernels, namely 3x3, 5x5, and 7x7. In multires block, Ibtehaz and Rahman [5] replace the 5x5 and 7x7 kernels with 3x3 kernels which have different filters and take the outputs from the three convolutional blocks and concatenate them together to extract the spatial features from different scales. In controlling the number of filters in a multires block, we need a parameter W which contains the coefficient value multiplied by the filter at each layer level. The filter values are 32, 64, 128, 256, 512 respectively at the respective layer levels. Determining the

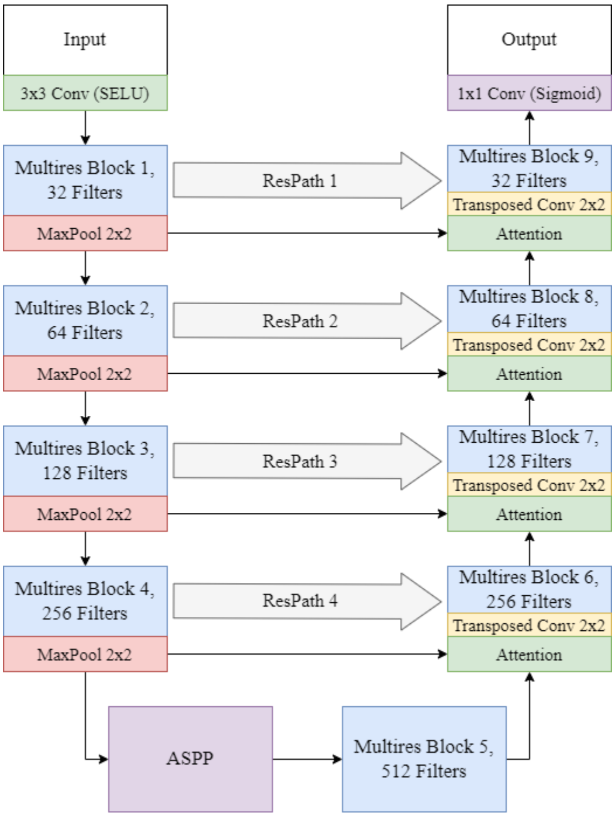
Input
3x3 Conv (SELU)
Output
1x1 Conv (Sigmoid)
Multires Block 1, 32 Filters
MaxPool 2x2
ResPath 1
Multires Block 9, 32 Filters
Transposed Conv 2x2
Attention
Multires Block 2, 64 Filters
MaxPool 2x2
ResPath 2
Multires Block 8, 64 Filters
Transposed Conv 2x2
Attention
Multires Block 3, 128 Filters
MaxPool 2x2
ResPath 3
Multires Block 7, 128 Filters
Transposed Conv 2x2
Attention
Multires Block 4, 256 Filters
MaxPool 2x2
ResPath 4
Multires Block 6, 256 Filters
Transposed Conv 2x2
Attention
ASPP
Multires Block 5, 512 Filters

Fig. 3. The architecture of the Proposed Model.

filter value for each kernel in multires is done by assigning $\frac{W}{6}$ to the first 3x3, $\frac{W}{3}$ to the second 3x3, and $\frac{W}{2}$ to the third 3x3. For example, multires block 1 has a filter value of 32. Thus, the filter values for each kernel on multires block 1 are 5.33 in the first 3x3, 10.67 in the second 3x3, and 16 in the third 3x3. Multires block use 1x1 convolutional layers which allow the model to comprehend some additional spatial information.

Fig. 4 is an illustration of the multires block structure. Where the multires block concatenates the output of three 3x3 convolution blocks with different filter values and also uses a residual connection with 1x1 convolution layer

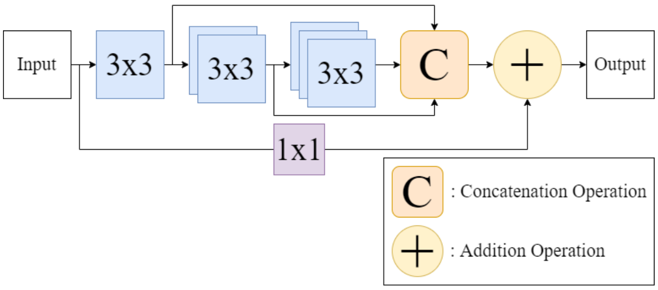

Fig. 4. Multires block structure illustration.

B. ResPath

The use of a skip connection between the encoder and decoder is considered to cause a semantic gap, Ibtehaz and Rahman [5] replaces the skip connection between the encoder and decoder with a residual path. Residual path is a combination of feature maps from encoder to decoder. The input from the encoder will pass through the convolution layer with residual connections. A 3x3 filter is used in the convolution layer and a 1x1 filter is used in the residual connection as illustrated in Fig. 5.

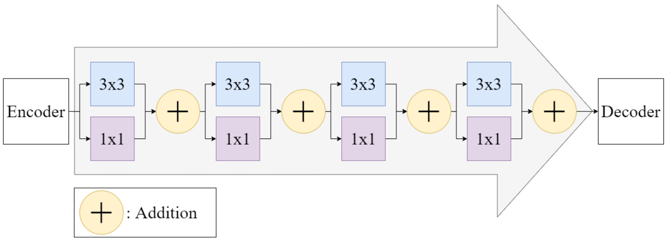

Fig. 5. Respath block structure illustration.

C. SELU Activation Function

The activation function used in MultiResUNet is ReLU (Rectified Linear Unit). One of the advantages of ReLU is the fast calculation speed because it does not contain complicated operations when the input is greater than zero. However, there is a potential problem when the input is less than or equal to zero because the output will be equal to zero and make the neuron not learn [9].

To solve this problem, this study uses SELU as an activation function, as shown in (1).

$$SELU(x) = \begin{cases} x & \text{if } x > 0 \\ \alpha e^x - \alpha & \text{if } x \leq 0 \end{cases} \tag{1}$$

while $\alpha \approx 1.6732632423543772848170429916717$ [10]

D. ASPP

The use of ASPP makes it possible to extract features from different scales. Atrous convolution allows controlling the field-of-view for capturing multi-scale information precisely. With the same set of dilation rates and functions as [6], in this study, ASPP was applied as part of the bridge between the encoder and decoder, where ASPP was used to overcome the problem of the different size of the systole and diastole segmentation objects by providing multi-scale information.

E. Attention

Attention mechanism is one of the important concepts in neural networks that has been applied to various domains and has been proven to improve model performance [11]. In segmentation problems using U-net, Jha et al. [6] use attention blocks on the decoder that is connected to the encoder so that the encoder can encode all the information from the polyp image into a vector with fixed dimensions. An important advantage of using the attention mechanism is that it can be used with a variety of input sizes and can improve model performance by allowing the model to focus on important areas of the feature map [6]. In this study, we add an attention block to the decoder connected to the encoder as in [6] so that the model is able to capture the left ventricle area more effectively.

III. Experiment

A. Dataset

The dataset used in this study is from the Echonet-Dynamic dataset. Echonet-Dynamic contains 10036 videos measuring 112x112 pixels and will be extracted into frames based on the available masks, which are then used for training and evaluation of the proposed model. The process of extracting video into frames based on available masks is based on research conducted by Ouyang et al. [3]. In general, Echonet-Dynamic provides data distribution for training, validation, and testing with 7465, 1289, and 1282 videos respectively. This study uses this setup without changing the composition of each data.

B. Training Process

The training model that has been built is carried out in the same way based on [3]. Frames and masks for training data are created based on the available masks.

The training process is carried out using Nvidia Tesla T4 with 15GB of memory. In general, the following hyperparameters are defined for training the proposed model:

- Epoch: 50
- Batch Size: 16
- Optimizer: Adam
- Learning Rate: 1e-3
- Loss Function: Binary Cross Entropy

In one iteration, in addition to calculating the loss value, the intersection and union between the prediction results and the mask are also calculated to get the dice coefficient value.

C. Evaluation

In conducting the evaluation, the resulting segmentation results will be measured using the calculation of the dice coefficient. The Dice Coefficient measures the overlap between the segmentation results and the mask or ground truth [12]. For example, x is the segmentation result area, and y is the mask area, then the Dice Coefficient measure is defined as follows:

$$DC(x, y) = \frac{2(x \cap y)}{x + y} \tag{2}$$

In measuring the performance of the model, it will calculate the value of twice the number of pixels in the intersecting image divided by the number of the two images. The greater the value of the dice coefficient, the closer the segmentation results to the mask or ground truth. The model will also be compared with MultiResUNet and DeeplabV3-Resnet50 in [3].

IV. Results and Discussion

In evaluating the results of this study, the measurement used to evaluate the results of segmentation is using the dice coefficient. Furthermore, the model that has been trained is tested on the test data and compared with the MultiResUNet and DeeplabV3-Resnet50 models. The MultiResUNet model was previously trained and tested on the Echonet-Dynamic dataset. While the test results on the DeeplabV3-Resnet50 model are taken based on [3]. Because the preprocessing of the data is done in the same way as [3], the test data used in the three models are exactly the same.

Table I shows the proposed model has been able to provide increased performance compared to the previous research models, namely MultiResUNet, by producing a dice coefficient value of 0.9206 in the overall test object compared to the MultiResUNet model which produces a dice coefficient value of 0.9066.

TABLE I
Comparison of proposed model test results on the overall data segmentation with MultiResUNet

Model Name	Dice Coefficient (Overall)
MultiResUNet	0.9066
Proposed Model	**0.9206**

Table II shows that the proposed model produces the best dice coefficient on the systole and diastole object tests compared to the MultiResUNet and Deeplabv3-Resnet50 models, which produces the value of 0.9045 on the systole object and 0.9306 on the diastole object. This is better than Deeplabv3-Resnet50 which produces the value of 0.903 on the systole object and 0.927 on the diastole object, also MultiResUNet which produces the value of 0.8865 on the systole object and 0.9193 on the diastole object.

TABLE II
Comparison of proposed model test results on systole and diastole segmentation with other models

Model Name	Dice Coefficient	
	Systole	***Diastole***
Deeplabv3-Resnet50	0.903	0.927
MultiResUNet	0.8865	0.9193
Proposed Model	**0.9045**	**0.9306**

In the test on the whole object, the value between the proposed model and the results on [3] cannot be compared because in this study there is no information about the results of the overall object test.

For other scenarios, we test the proposed model by removing ResPath or changing ResPath to default skip connection. Table III shows the dice coefficient value of the proposed model with ResPath and the proposed model without ResPath. The proposed model with ResPath shows better results. This is because ResPath is able to reduce the unbalanced state or difference between the encoder and decoder in the model.

Figure 6 shows the segmentation results of the three models. We can see in Figure 6(e) that the proposed model produces

TABLE III
COMPARISON OF TEST RESULTS BETWEEN PROPOSED MODEL WITH RESPATH AND PROPOSED MODEL WITHOUT RESPATH

Model Name	**Dice Coefficient (Overall)**
Without ResPath	0.9117
With ResPath	**0.9206**

results that are more precise to ground truth than other models, with the Deeplabv3-Resnet50 model tends to produce smaller results.

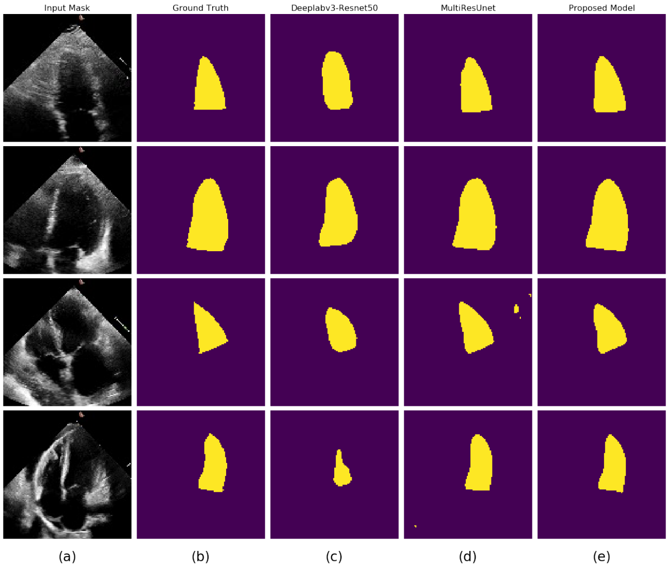

Fig. 6. Qualitative results of the three models using the Echonet-Dynamic dataset, (a) Input image, (b) Ground Truth, (c) Deeplabv3-Resnet50 model results, (d) MultiResUnet model results, (e) Proposed model results

V. CONCLUSION

In this study, we modified the MultiResUNet architecture by adding an ASPP block and an attention block for more accurate segmentation of left ventricular in echocardiogram. Models were trained and evaluated using the echonet-dynamic dataset. The segmentation results are then evaluated using the Dice Similarity Coefficient and obtain an accuracy value of 0.9206. These results were then compared with the Multiresunet and Deeplabv3-Resnet50 models using the same dataset. From the comparison results, in general, the proposed model outperforms the Multiresunet and Deeplabv3-Resnet50 models.

ACKNOWLEDGMENT

We gratefully acknowledge the support from Tokopedia-UI AI Center, Faculty of Computer Science, University of Indonesia, for the used of NVIDIA DGX-1 for running the experiments.

REFERENCES

[1] K. Deng, Y. Meng, D. Gao, J. Bridge, Y. Shen, G. Lip, Y. Zhao, and Y. Zheng, “TransBridge: A lightweight transformer for left ventricle segmentation in echocardiography,” *Simplifying Medical Ultrasound*, pp. 63–72, 2021.

[2] D. Ouyang, B. He, A. Ghorbani, M. P. Lungren, E. A. Ashley, D. H. Liang, and J. Y. Zou, “Echonet-dynamic: a large new cardiac motion video data resource for medical machine learning,” *NeurIPS ML4H Workshop: Vancouver, BC, Canada*, 2019.

[3] D. Ouyang, B. He, A. Ghorbani, N. Yuan, J. Ebinger, C. P. Langlotz, P. A. Heidenreich, R. A. Harrington, D. H. Liang, E. A. Ashley, and J. Y. Zou, “Video-based AI for beat-to-beat assessment of cardiac function,” *Nature*, vol. 580, no. 7802, pp. 252–256, 2020.

[4] N. Siddique, S. Paheding, C. P. Elkin, and V. Devabhaktuni, “U-Net and its variants for Medical Image Segmentation: A review of theory and applications,” *IEEE Access*, vol. 9, pp. 82031–82057, 2021.

[5] N. Ibtehaz and M. S. Rahman, “MultiResUNet : Rethinking the U-net architecture for multimodal biomedical image segmentation,” *Neural Networks*, vol. 121, pp. 74–87, 2020.

[6] D. Jha, P. H. Smedsrud, M. A. Riegler, D. Johansen, T. D. Lange, P. Halvorsen, and H. D. Johansen, “ResUNet++: An Advanced Architecture for Medical Image Segmentation,” *2019 IEEE International Symposium on Multimedia (ISM)*, 2019.

[7] A. Amer, X. Ye, M. Zolgharni, and F. Janan, “ResDUnet: Residual dilated unet for left ventricle segmentation from echocardiographic images,” *2020 42nd Annual International Conference of the IEEE Engineering in Medicine*; Biology Society (EMBC), 2020.

[8] Y. Xu, T. Xiao, J. Zhang, K. Yang, and Z. Zhang, ”Scale-invariant convolutional neural networks,” *arXiv preprint arXiv:1411.6369 [cs.CV]*, 2014.

[9] Q. Yang, Y. Li, M. Zhang, T. Wang, F. Yan, and C. Xie, “Automatic segmentation of COVID-19 CT images using improved multiresunet,” *2020 Chinese Automation Congress (CAC)*, 2020.

[10] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, ”Self-Normalizing Neural Networks,” *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[11] Z. Niu, G. Zhong, and H. Yu, ”A review on the attention mechanism of deep learning,” *Neurocomputing*, vol. 452, pp. 48-62, 2021.

[12] L. R. Dice, “Measures of the amount of ecologic association between species,” *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

978-1-6654-8950-8/22/$31.00 ©2022 IEEE