Jurnal Ilmu Komputer dan Informasi (Journal of Computer Science and Information)
16/2 (2023), 163-169. DOI: http://dx.doi.org/10.21609/jiki.v16i2.1165

Encoder-Decoder with Atrous Spatial Pyramid Pooling for Left Ventricle Segmentation in Echocardiography

Fityan Azizi[1], Mgs M Luthfi Ramadhan[1], Wisnu Jatmiko[1]

[1] Faculty of Computer Science, Universitas Indonesia, Depok, 16424, Indonesia

Email:fityan.azizi@ui.ac.id

Abstract

Assessment of cardiac function using echocardiography is an essential and widely used method. Assessment by manually labeling the left ventricle area can generally be time-consuming, error-prone, and has inter-observer variability. Thus, automatic delineation of the left ventricle area is necessary so that the assessment can be carried out effectively and efficiently. In this study, encoder-decoder based deep learning model for left ventricle segmentation in echocardiography was developed using the effective CNN U-Net encoder and combined with the deeplabv3+ decoder which has efficient performance and is able to produce sharper and more accurate segmentation results. Furthermore, the Atrous Spatial Pyramid Pooling module were added to the encoder to improve feature extraction. Tested on the Echonet-Dynamic dataset, the proposed model gives better results than the U-Net, DeeplabV3+, and DeeplabV3 models by producing a dice similarity coefficient of 92.87%. The experimental results show that combining the U-Net encoder and DeeplabV3+ decoder is able to provide increased performance compared to previous studies.

Keywords: *Echocardiography, Left Ventricle Segmentation, Deep Learning, DeeplabV3+, U-Net*

1. Introduction

Assessment of cardiac function on echocardiography is generally done by observing the value of the left ventricle ejection fraction ratio (LVEF). Manual and semi-automatic assessment requires delineation of left ventricle area at both the end of contraction (end systole) and the end of relaxation (end diastole) to obtain the value of the LVEF ratio [1]. Manual delineation of the left ventricular area can be time-consuming because it needs to be done carefully and is inconsistent due to high inter-observer variance [2–4]. Thus, automatic delineation of the left ventricular area is necessary so that the assessment of cardiac function on echocardiography can be carried out effectively and efficiently.

Automated depiction of the left ventricle area can be performed by segmentation using deep learning. Models with the type of encoder-decoder are models that are widely used and many provide good performance results in image segmentation work [5–8]. U-Net is a Deep Learning model with an encoder and decoder form that is specifically designed to carry out image segmentation work, especially in the biomedical field. The encoder part consists of the composition of the Convolutional Neural Network (CNN) which provides information for classification. The encoder in U-Net is claimed to be able to extracted features efficiently. The decoder contains the combination of features of the encoder through the skip connections and the expansion of the image size to restore the original size, so that the image segmentation is formed [6]. U-Net and modification have been widely used in various biomedical images such as liver, polyp, and pancreas [9–12]. Although it has been widely applied to various types of medical data and has an efficient encoder, some modifications to improve U-net performance are done by maintaining the encoder and changing the decoder including the skip connections[13–16]. This is because the combination of features between encoder and decoder through skip connection can actually limit the performance of the model, Because there is no guarantee that even though the size of the features between encoder and decoder is the same, the two features that are combined have the same

semantics between the two [13].

Another model with Encoder-Decoder architecture is Deeplabv3+. Chen et al. [7] develops a model from the previous version, namely Deeplabv3 [17] by providing a simple but efficient decoder so that it can produce sharper and more accurate segmentation. One of the advantages of Deeplabv3 is the use of Atrous Spatial Pyramid Pooling (ASPP) that is able to extract features on different scales.

In supporting research related to the left ventricle segmentation in echocardiography, Ouyang et al. [4] provides dataset of echocardiographic image called Echonet-Dynamic, which contains 10030 echocardiographic videos and anotation of segmentation objects carried out by experts. Not only providing dataset, Ouyang et al. [4] also conducts experiments by segmentation of the left ventricle using the Deeplabv3 model with a resnet-50 backbone. As a result, by using the Dice Similiarity Coefficient evaluation measurement, the model produced a value of 95%. The Dice Similiarity Coefficient calculation is done by measuring overlapping between the results of segmentation and the Ground Truth.

In this study, a Deep Learning model for the left ventricle segmentation in echocardiography will be developed by combining the encoder from U-Net with the ASPP and Decoder modules from Deeplabv3+. The ASPP module will be used to improve the ability of the model in extraction of features on objects that vary in the left ventricle when the systole and diastole. U-Net encoder is used to extract features efficiently and used decoder from Deeplabv3+ which has a simple but efficient design to avoid semantic gap problems, improve model efficiency, and also has the ability to produce more sharper and accurate segmentation. The main contributions of this study are summarized as follows:

- Developed a model that is more efficient than the original model used as encoder and decoder in this study, U-Net and DeeplabV3+ by having fewer parameters.
- Compared the proposed model with the U-Net, DeeplabV3+ and DeeplabV3 models from [4].

2. Proposed Model

The model architecture used in this study is in the encoder-decoder form with the encoder used to extract features and the decoder used to produce segmentation results. The encoder used is a CNN encoder taken from U-Net and claimed to be able to extract features efficiently. At the end of the encoder, ASPP is added so that the model is able to extract features at different scales. This is useful for recognizing objects well during systole and diastole because the two objects have different sizes.

Figure 1 is an illustration of the U-Net architecture, where there are skip connections between low level features from encoder to decoder. Furthermore,

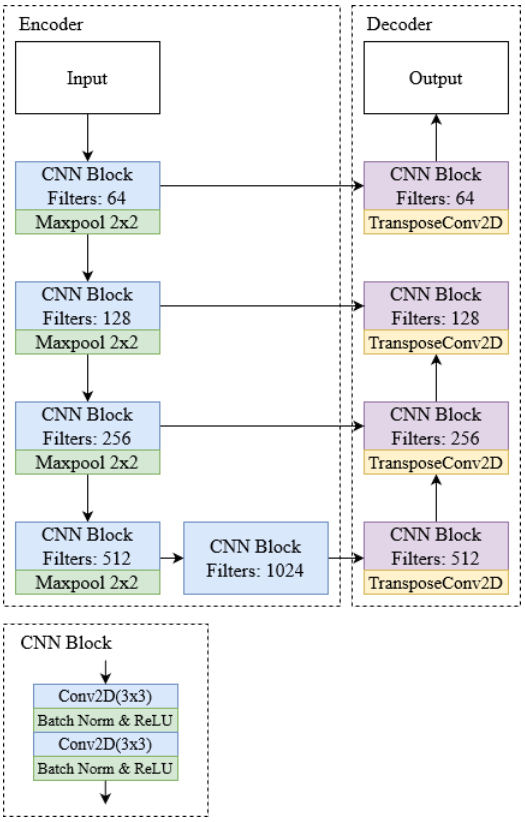

Figure 1. The architecture of U-Net model (reproduced).

the model architecture in this study does not use the decoder from U-Net. This is because in U-Net, combining features between encoder and decoder via skip connection can limit model performance because there is semantic gap between encoder and decoder, also there is no guarantee that without using a skip connection can improve the model performance. So the model architecture in this study uses a decoder from DeeplabV3+ which can produce sharper and more accurate segmentation.

2.1. Encoder

The encoder used in the proposed model is the U-Net encoder, which contains five CNN blocks. In one CNN block, it contains two 3x3 convolutions followed by ReLU activation function. At the end of the block there is a 2x2 Max Pooling operation for downsampling (reducing image resolution). The

number of channels in the first block on the encoder has a value of 64, then the number of channels will double each subsequent block [6].

U-Net encoder is claimed to be able to efficiently extract high-level features. Furthermore, the encoder in the proposed model will utilize the ASPP module from DeeplabV3+ [7]. ASPP can capture context information at different scales. The use of this module is used with the aim that the model can work well on systole and diastole objects that have various sizes. The ASPP module contains atrous filters with different sizes. The rates used are 6, 12, and 18.

Figure 2 is an illustration of the encoder used in this study. The leftmost of the figure is the CNN block. The rightmost of the figure is the encoder block module with a filter size of 64 at the initial block, then doubling in each subsequent block. the ASPP module is added at the end of the encoder.

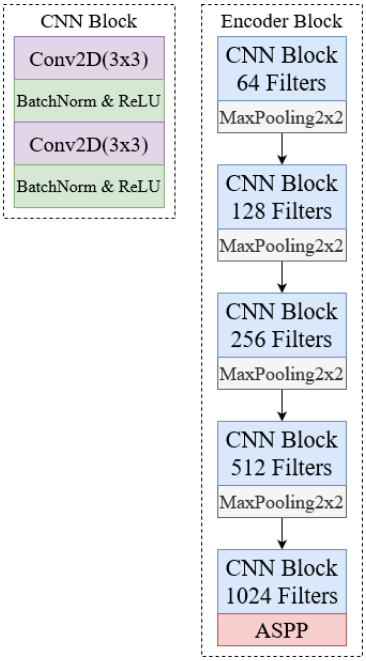

Figure 2. Encoder Block illustration.

2.2. ASPP

Atrous Spatial Pyramid Pooling (ASPP) has the ability to extract and provide multi-scale information [7, 17]. ASPP uses three 3x3 dilated convolutions, one 1x1 convolution, and global average pooling in parallel. Each convolution has a different dilation ratio, followed by batch normalization and activation of the ReLU function. Next, the four parallel convolutions and global average pooling are combined, finally the 1x1 convolution is carried out. Figure 3 is an illustration of the ASPP module, where the dilation ratio is 6, 12, and 18. In this study, ASPP

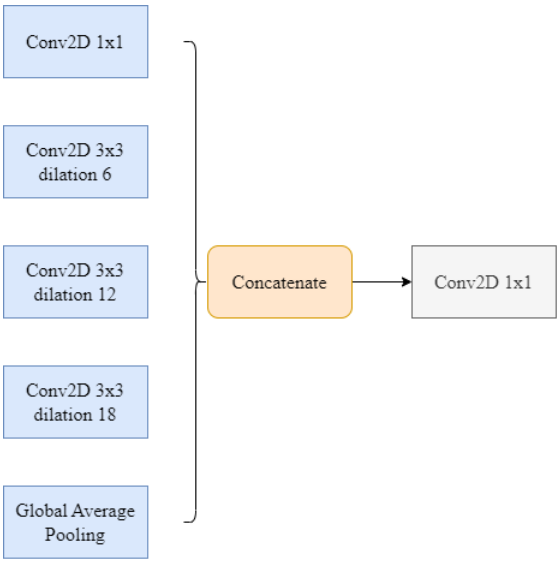

Figure 3. Atrous Spatial Pyramid Pooling (ASPP) Block illustration.

is used at the end of the encoder to improve the encoder in extracting features by providing multi-scale information.

2.3. Decoder

The decoder used in this study is a decoder from DeeplabV3+. There are two features from the encoder that enter the decoder, namely the high level feature from the end of the encoder and the low level feature from the start of the encoder. Furthermore, the number of channels that enter the decoder is adjusted by 1x1 convolution. The high-level features are then upsampled by bilinear interpolation with a factor of 4 and combined with the low-level features. Finally, the features that have been combined are then carried out a 3x3 convolution and upsample by performing a bilinear interpolation with a factor of 4 to get the segmentation results [7]. In general, the proposed model architecture is illustrated in Figure 4. The low level feature given to the decoder is the second block of the encoder, which has a total of 128 channels and is then adjusted or reduced by 1x1 convolution to 48.

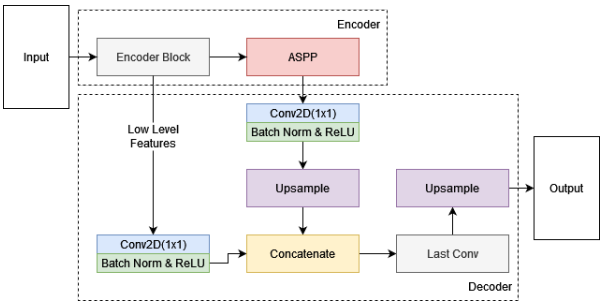

Figure 4. The architecture of the Proposed Model.

3. Experiments

3.1. Dataset

The dataset used for the experiments in this study is the Echonet-Dynamic dataset [4]. This dataset contains 10030 videos with each video having a height and width of 112x112. In order to be used in left ventricle segmentation, frame extraction was performed for each end systole and end diastole based on annotations of 20048 images. From the available images, the data is separated into 3 parts with 14920 images used for training, 2576 images for validation, and 2552 images for testing. This study uses this setup without changing the composition of each data.

This study also uses another public dataset that is used to test the model, namely the CAMUS dataset.[3]. This dataset contains 2-dimensional echocardiography images of apical two-chamber view (A2C) and apical four-chamber view (A4C) taken from 500 patients. In this study, this dataset is used to establish the versatility of the proposed model because it has various image qualities, from good quality images with a clear left ventricle to poor quality images with an unclear and blurry left ventricle[18].

3.2. Training Process

The training model that has been built is carried out in the same way based on [4]. There are only differences in optimizer, batch size, and learning rate. The training process is carried out using Nvidia 1650 with 4GB of memory. In general, the following hyperparameters are defined for training the proposed model: In this study, no pretrained weights were used in the encoder.

Table 1. Hyperparameters used in Training scenario.

Parameter	Value
Epoch	50
Optimizer	Adam
Batch Size	8
Learning Rate	1e-4
Loss Function	Binary Cross Entropy

3.3. Evaluation

The evaluation metric used to measure model performance is the dice similarity coefficient (DSC). This measurement is carried out by calculating the overlapping pixels between the segmentation results and the ground truth annotation. DSC measurement is defined as follows:

$$DSC(x, y) = \frac{2(x \cap y)}{x + y} \tag{1}$$

where x is the segmentation result and y is the ground truth annotation, the equation calculates the value of twice the number of intersecting pixels divided by the number of the two images. Values in DSC are in the range 0 to 1, where as the DSC value gets closer to 1, the more similar the segmentation results are to the ground truth annotation [19].

4. Results and Discussion

In this study evaluation were carried out by comparing the DSC values performed by the models. The data used in training, validation, and testing is used based on the train-validation-test split. The composition of the train-validation-test data on the Echonet-Dynamic dataset is defined by default[4]. So that model performance can be compared with previous studies in an easy and fair way. This study uses a performance test scenario on test data with the same composition as set by default and was also used in previous studies, so the train-validation-test data used is the same. Using other scenarios (e.g. cross-validation) can result in unfair performance comparisons with other studies.

The performance of the proposed model is compared with the U-Net, DeeplabV3+, and DeeplabV3 models from [4]. The evaluation was carried out by comparing the DSC values at end systole, end diastole, and the overall values of both. The U-Net and DeeplabV3+ models were previously trained and tested on the Echonet-Dynamic dataset and the backbone used on DeeplabV3+ is Resnet101. Test results on DeeplabV3 are taken based on [4]. Overall, data pre-processing is done in the same way as [4], so

that the data used during training and testing are all the same.

Table 2 shows the DSC values produced by the four models in the Echonet-Dynamic test data. In end systole object test, the proposed model produces the best value with a value of 0.9158. The U-Net, DeeplabV3+, and DeeplabV3 models produce DSC values of 0.9133, 0.9080, and 0.903 respectively. In end diastole object test, the proposed model is also able to give the best value with a value of 0.9367. In descending order, the U-Net, DeeplabV3+, and DeeplabV3 models from [4] produce DSC values of 0.9343, 0.9292, and 0.927 respectively. Furthermore, the proposed model produces the highest DSC value compared to the other three models with an overall value of 0.9287. The U-Net model produces 0.9262 values, the DeeplabV3+ model generates 0.9210 values, and the lowest is DeeplabV3 from [4] with 0.915 values. This shows that the proposed model using encoder from U-Net and decoder from DeeplabV3+ is able to provide increased performance compared to the two models.

Table 2. Comparison of segmentation results between the proposed model and the DeeplabV3, DeeplabV3+, and U-Net models. The results from DeeplabV3 are taken based on [4].

Model	Dice Similiarity Coefficient		
	Systole	Diastole	Overall
DeeplabV3[4]	0.903	0.927	0.915
DeeplabV3+	0.9080	0.9292	0.9210
U-Net	0.9133	0.9343	0.9262
Proposed Model	**0.9158**	**0.9367**	**0.9287**

Table 3 displays the total number of parameters used by the four models. Especially when compared to the U-Net and DeeplabV3+ models with 39.39M and 59.33M parameters respectively, the proposed model has 28.01M parameters. This shows that the proposed model is able to work more efficiently because it provides better performance results and uses parameters which is less than the U-Net and DeeplabV3+ models. The proposed model also has the lowest number of parameters compared to the other three models because the DeeplabV3-Resnet50 model has 39.6M parameters.

Table 3. Comparison of the number of parameters between the proposed model and the DeeplabV3+ and U-Net models.

Model	Number of Parameters (Million)
DeeplabV3	39.6M
DeeplabV3+	59.33M
U-Net	39.39M
Proposed Model	**28.01M**

Table 4 shows the running time when doing inference or segmentation for one image. The proposed model performs segmentation with the shortest time compared to unet and deeplabv3+. The proposed model requires 0.00227 seconds to perform segmentation for one image, while the U-Net model requires 0.00256 seconds, the DeeplabV3 model from [4] requires 0.00256 seconds and the DeepLabV3+ model requires 0.00907 seconds. this shows the effectiveness of the model in terms of model size and the time needed to perform segmentation.

Table 4. Comparison of the running time of segmentation for one image between the proposed model and the DeeplabV3+ and U-Net models.

Model	Running Time (Second)
DeeplabV3[4]	0.00596
DeeplabV3+	0.00907
U-Net	0.00256
Proposed Model	**0.00227**

Figure 5 is an example of segmentation results from the proposed model, DeeplabV3+, and U-Net. Figure (a) shows the input or the original image, (b) are the ground truths, and (c) to (e) are the segmentation outputs of the DeeplabV3+, U-Net, and proposed model respectively. Based on the four results shown, the proposed model can produce segmentation similar to ground truth and better than the results from DeeplabV3+ and U-Net.

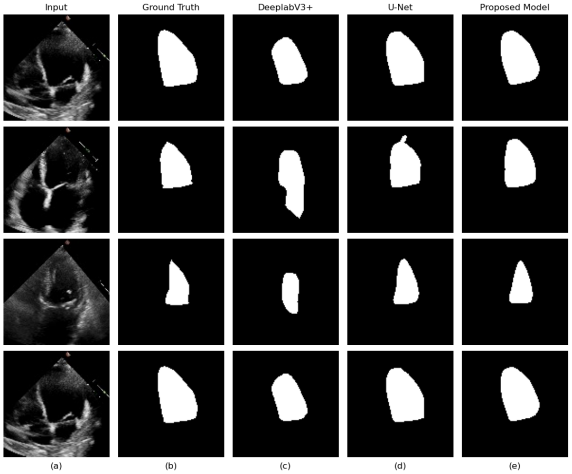

Figure 5. Segmentation results from the DeeplabV3+, U-Net, and proposed models. (a) image input, (b) ground truth, and (c) to (e) are the output segmentation from DeeplabV3+, U-Net, and the proposed model.

Table 5 is the test result on the CAMUS dataset. In contrast to the Echonet-Dynamic dataset, this

Table 5. Comparison of segmentation results between the proposed model and the DeeplabV3 from [4], DeeplabV3+, and U-Net models on the CAMUS dataset.

Model	Dice Similiarity Coefficient		
	Systole	Diastole	Overall
DeeplabV3[4]	0.8781	0.9081	0.8963
DeeplabV3+	0.8856	0.9164	0.9041
U-Net	0.9037	0.9257	0.9170
Proposed Model	0.9006	0.9207	0.9128

dataset has various image quality from good quality images with a clearly visible left ventricle to poor quality images with an unclear and blurry left ventricle. The proposed model is able to produce a DSC value of 0.91287 for the overall test data. With the end systole object test produces a value of 0.9006 and the end diastole object test produces a value of 0.9207. These results outperform the DeeplabV3+ model performance which produces a DSC value on the overall test data of 0.9041 and DeeplabV3 from [4] with a value of 0.8963, but produces a slightly lower value than U-Net with a DSC value of 0.9170. This shows that although the proposed model is capable of providing good results, it cannot outperform the performance results of the U-Net model on CAMUS datasets which have various image qualities.

5. Conclusion

In this study, an encoder-decoder model was developed for left ventricle segmentation in echocardiography using encoder from U-Net and decoder from DeeplabV3+ and adding ASPP to the encoder. The proposed model is then trained and tested using the Echonet-Dynamic dataset and produces a Dice Similiarity Coefficient value of 0.9287. These results were compared with the U-Net, DeeplabV3+, and DeeplabV3 models. In general, the proposed model proves to be more efficient than the other three models because it produces the best performance with the least number of parameters. The proposed model can also be used as a tool for cardiac function assessment.

References

[1] A. Ghorbani, D. Ouyang, A. Abid, B. He, J. H. Chen, R. A. Harrington, D. H. Liang, E. A. Ashley, and J. Y. Zou, "Deep learning interpretation of echocardiograms," *npj Digital Medicine*, vol. 3, no. 1, p. 10, Jan 2020. [Online]. Available: https://doi.org/10.1038/s41746-019-0216-8

[2] R. M. Lang, L. P. Badano, V. Mor-Avi, J. Afilalo, A. Armstrong, L. Ernande, F. A. Flachskampf, E. Foster, S. A. Goldstein, T. Kuznetsova, P. Lancellotti, D. Muraru, M. H. Picard, E. R. Rietzschel, L. Rudski, K. T. Spencer, W. Tsang, and J.-U. Voigt, "Recommendations for cardiac chamber quantification by echocardiography in adults: An update from the american society of echocardiography and the european association of cardiovascular imaging," *Journal of the American Society of Echocardiography*, vol. 28, no. 1, pp. 1–39.e14, 2015.

[3] S. Leclerc, E. Smistad, J. Pedrosa, A. Østvik, F. Cervenansky, F. Espinosa, T. Espeland, E. A. R. Berg, P.-M. Jodoin, T. Grenier, C. Lartizien, J. D'hooge, L. Lovstakken, and O. Bernard, "Deep learning for segmentation using an open large-scale dataset in 2d echocardiography," *IEEE Transactions on Medical Imaging*, vol. 38, no. 9, pp. 2198–2210, 2019.

[4] 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.

[5] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," 2016.

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Springer International Publishing, 2015, pp. 234–241.

[7] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[8] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)*, 2016, pp. 565–571.

[9] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention u-net: Learning where to look for the pancreas," in *Medical Imaging with Deep Learning*, 2018. [Online]. Available: https://openreview.net/forum?id=Skft7cijM

[10] 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.

[11] D. Jha, P. H. Smedsrud, M. A. Riegler, D. Johansen, T. de Lange, P. Halvorsen, and H. D. Johansen, "Resunet++: An advanced architecture for medical image segmentation," 2019.

[12] 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. 82 031–82 057, 2021.

[13] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Transactions on Medical Imaging*, vol. 39, no. 6, pp. 1856–1867, 2020.

[14] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, and J. Wu, "Unet 3+: A full-scale connected unet for medical image segmentation," 2020.

[15] 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," in *Simplifying Medical Ultrasound*, J. A. Noble, S. Aylward, A. Grimwood, Z. Min, S.-L. Lee, and Y. Hu, Eds. Cham: Springer International Publishing, 2021, pp. 63–72.

[16] H. Wang, P. Cao, J. Wang, and O. R. Zaiane, "Uctransnet: Rethinking the skip connections in u-net from a channel-wise perspective with transformer," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, pp. 2441–2449, Jun. 2022.

[17] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *CoRR*, vol. abs/1706.05587, 2017. [Online]. Available: http://arxiv.org/abs/1706.05587

[18] A. Amer, X. Ye, and F. Janan, "Resdunet: A deep learning-based left ventricle segmentation method for echocardiography," *IEEE Access*, vol. 9, pp. 159 755–159 763, 2021.

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