DeepEar: A Deep Convolutional Network without Deformation for Ear Segmentation

Yuhan Chen [1], Wende Ke [1,*], Qingfeng Li [2], Dongxin Lu [2], Yan Bai [1], and Zhen Wang [1]

[1] Department of Mechanical and Energy Engineering, Southern University of Science and Technology, Shenzhen, China

[2] Mobile Health Management System Engineering Center, Hangzhou Normal University, Hangzhou, China

*Correspondence: kewd@sustech.edu.cn (W.K.)

***Abstract*—With the cross-application of robotics in various fields, machine vision has gradually received attention. As an important part in machine vision, image segmentation has been widely applied especially in biomedical image segmentation, and many algorithms in image segmentation have been proposed in recent years. Nowadays, traditional Chinese medicine gradually received attention and ear diagnosis plays an important role in traditional Chinese medicine, the demand for automation in ear diagnosis becomes gradually intense. This paper proposed a deep convolution network for ear segmentation (DeepEar), which combined spatial pyramid block and the encoder-decoder architecture, besides, atrous convolutional layers are applied throughout the network. Noteworthy, the output ear image from DeepEar has the same size as input images. Experiments shows that this paper proposed DeepEar has great capability in ear segmentation and obtained complete ear with less excess region. Segmentation results from the proposed network obtained Accuracy = 0.9915, Precision = 0.9762, Recal l= 9.9723, Harmonic measure = 0.9738 and Specificity = 0.9955, which performed much better than other Convolution Neural Network (CNN)-based methods in quantitative evaluation. Besides, this paper proposed network basically completed ear-armor segmentation, further validated the capability of the proposed network.**

***Keywords*—deep convolutional network, certain output size, ear segmentation**

I. Introduction

In recent years, robotics are constantly evolving and have been widely applied in industry [1, 2]. Besides, the rapid progress of robotic technique promotes the development healthcare engineering [3]. As the outbreak of novel coronavirus pneumonia (COVID-19) [4], robots are also applied to fight the epidemic [5, 6]. It is obvious that machine vision is attended greatly in automatic system [7]. Image segmentation plays a significant role in machine vision, researchers tend to create more efficient and accurate image segmentation methods.

Traditional image segmentation methods can be divided into two categories. One category of methods are graph-based image segmentation methods, the other are clustering-based segmentation methods. Graph-based image segmentation methods map images into weighted undirected graphs according to graph-theory, treated pixels as nodes, considered the image segmentation problem as a vertex partition problem of the graph, and obtained the best segmentation according to the minimum cutting criterion [8]. Researchers modified cutting criteria in order to improve the accuracy of image segmentation, representative methods are normalized cut method [9], graph cut method [10] and grab cut method [11]. However, graph-based methods still need an interaction with users when dealing with complex texture features of images. Many researchers concentrated on clustering in machine learning to solve image segmentation problems. The clustering methods are implemented through iteration, pixels similar in color, brightness, texture and other features are clustered to the same superpixel, so as to obtain the final image segmentation results. Compared to the original K-means algorithm [12], spectral clustering [13] introduced graph into clustering, obtained more accurate optimal solutions. According to mean-shift clustering, pixels with the same modulus point are segmented to the same region [14]. Another clustering method called Simple Linear Iterative Clustering (SLIC), from which color images were transformed into 5-dimensional feature vectors in Lab color space as well as XY coordinates [15]. Though clustering-based methods obtained good segmentation results, time complexity and spatial complexity of these algorithms are too large to really be really put into application.

With the development of computer and artificial intelligence, deep learning has gradually been applied to image segmentation technology, which brings many breakthroughs to the field of image segmentation. Long et al. proposed Fully Convolutional Networks (FCN) [16] which pioneered deep learning to solve the image segmentation problem, FCN included a forward inference process and a backward learning process and converted coarse segmentation results to dense results. Similar to FCN, SegNet was proposed with a combination of encoder and decoder [17]. Olaf *et al.* [18] proposed U-Net, which is now popular for segmentation tasks in all fields, U-Net

Manuscript received February 7, 2023; revised April 11, 2023; accepted July 20, 2023.

has a symmetric architecture of down sampling as well as up sampling and connected tensors in down sampling as well as up sampling of the same scale, which combined semantic and positional information together. Later on, Zhao *et al.* [19] proposed PSPNet, which used residual network (ResNet) as a feature extractor and extract different pattern from the input image using an expansion network, this network is characterized by the application of pyramid pooling module for multi-scale convolution. Having concentrated on atrous convolution, Chen *et al.* [20, 21] proposed a serious of DeepLab networks for image segmentation, and they constantly improved the network architecture of DeepLab [22, 23], which has also achieved good segmentation results. Recently, researchers tend to introduce Generative Adversarial Network (GAN) into image segmentation and obtained pretty results [24, 25]. However, the texture features of human ears are pretty complex, and existing few datasets, realizing the identification and segmentation of ear from images with simply RGB information is undoubtedly a challenge [26].

Contributions in this paper are summarized in the following:

- **A convolutional network for segmentation:** Inspired by DeepLab v3+, we constructed an ear segmentation network called DeepEar combines spatial pyramid pooling and the encoder-decoder architecture, in view of the capability of convolution layers for encoder, we continue to apply the convolution layers with stride 2 instead of pooling layers and atrous convolution for expanding the receptive field of pixels.
- **A pixel segmentation for ear from RGB images:** We provide a novel method for ear segmentation without deformation, output ear from the proposed method has the same size as input images. Dilation rates in spatial pyramid block from the proposed network are adjusted for the particular texture features of human ears. In addition, architecture of the proposed network is simplified to satisfy the actual demands of the automated medical machines.

II. Network Architecture

In this section, we will discuss architecture of the proposed DeepEar. First, we will introduce the overall architecture of the proposed network. Later the encoder-decoder architecture layer details will be revealed. Finally, we will explain the spatial pyramid block.

A. Global Architecture

This paper proposed DeepEar is a combination of encoder-decoder architecture and spatial pyramid block, the global architecture of DeepEar as is shown in Fig. 1. Size of output tensors from Encoder 1 and Encoder 2 are both shrank four times and size of output tensors from Decoder 1 and Decoder 2 are both magnified four times. Besides, DeepEar contains two concat connections, distinct from DeepLab v3+, a concat connection is raised between tensors before the first encoding and tensors after the last decoding, the comparison of encoder-decoder architecture of DeepEar as well as U-Net and DeepLab v3+ as shown in Fig. 2. It is clear that DeepEar ensures the output image with more accurate pixel information while simplifying the network structure.

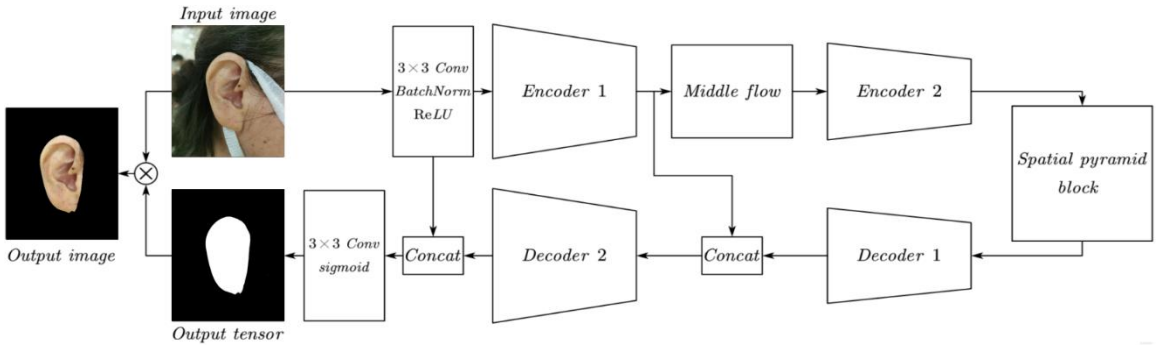


Figure 1. Global architecture of DeepEar.

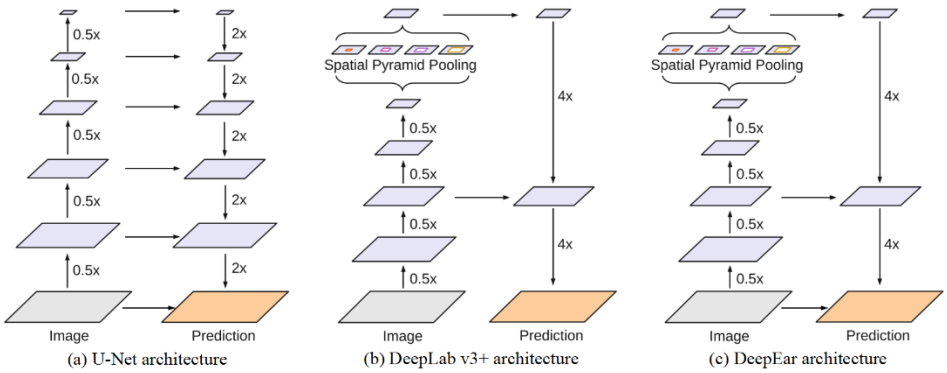
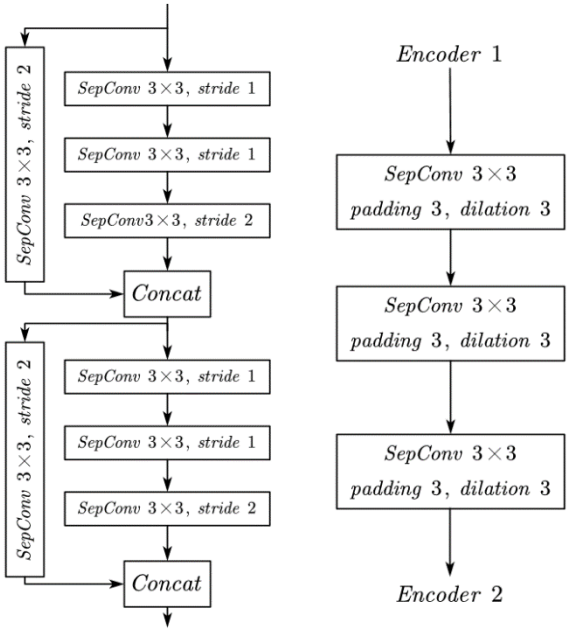


Figure 2. Network architecture comparison of U-Net, DeepLab and DeepEar.

The output tensor is a single channel image with the same shape as input image (a binary map where target area is true otherwise false). The output segmentation results are obtained by element-wise production of output tensors and input images, which not only ensures the image segmentation effect and retains original detailed features of raw ear images.

B. Encoder-Decoder Layer Details

The encoding process can be divided into three parts: Encoder 1, Middle flow and Encoder 2. Architecture of Eecoders and Middle flow as shown in Fig. 3, all SepConv layers including a convolutional layer, a batch norm layer and a ReLU layer. The Encoder consists of a backbone and two auxiliary layers for increasing the encoding influence from raw information as is shown in Fig. 3(a) and the two encoders are in the same architecture. Besides, both Encoder1 and Encoder2 apply convolutional layers with stride 2, the specific convolutional layers have similar function to pooling layers and they have the capability for training. Middle flow is a block of three SepConv layers in series as is shown in Fig. 3(b). The atrous convolution with dilation = 3 is applied for wider receptive field and padding = 3 is set to prevent deformation.



(a) Encoder1 & Enaoder2 (b) Middle flow

Figure 3. Details of encoding process.

The decoding process achieved by two simple four-time-bilinear upsample and there is a SepConv layer in the front of each Decoder, the SepConv layer in Decoder 1 is to fuse concat information from spatial pyramid block and the SepConv layer in Decoder 2 is to fuse processed information as well as the raw information.

C. Spatial Pyramid Block

Spatial pyramid block constructs a parallel connection and process encoded tensors separately from five parts. The five parts are 1×1 convolution, convolution with dilation 3, convolution with dilation 6, convolution with dilation 9, and a pooling part. To cater to the size of the ear image, dilation rate of the atrous convolutional layers are adjusted to 3, 6 and 9. The atrous convolutional layers as well as 1×1 convolution and pooling implement multi scale semantic segmentation, results of above five parts are adequately fused according to concat and sent to Decoders, the architecture of spatial pyramid block as shown in Fig. 4.

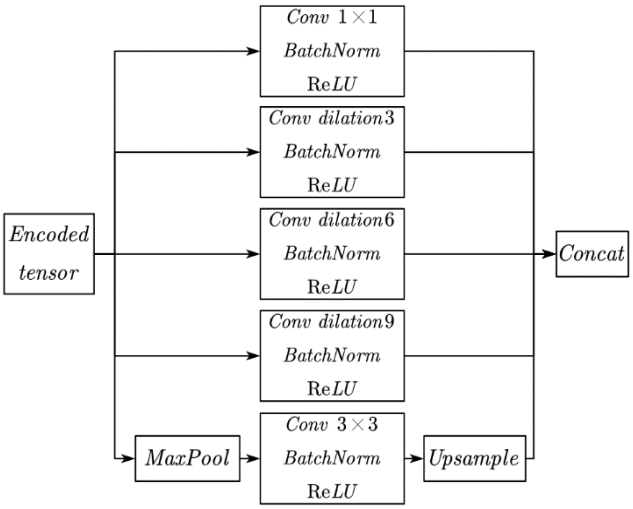


Figure 4. Architecture of spatial pyramid block.

III. Dataset Collection

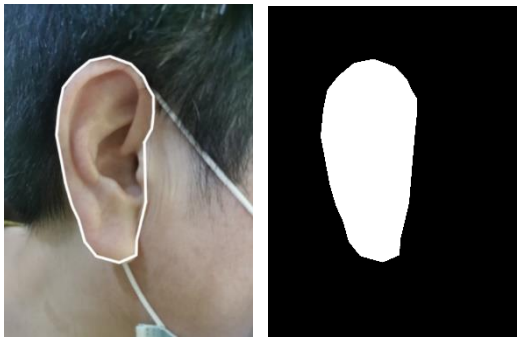
A. Dataset Source

In order to improve enhance the generalization of the segmentation networks, the dataset with two sources collected 168 ear images are applied in DeepEar training and testing. One is 68 standardized images collected by the hospital through professional equipment, the image size is standardized at 500 pixel × 500 pixel, the other is 100 unstandardized images taken from Mishixiang in Hangzhou through the mobile phone, the length-width ratio of the images are 4:3 [27]. 60 images from the standardized images and 90 images from unstandardized images are randomly chosen to obtain raw images in training set and the rest 18 images are treated as raw images in testing set.

B. Dataset Processing

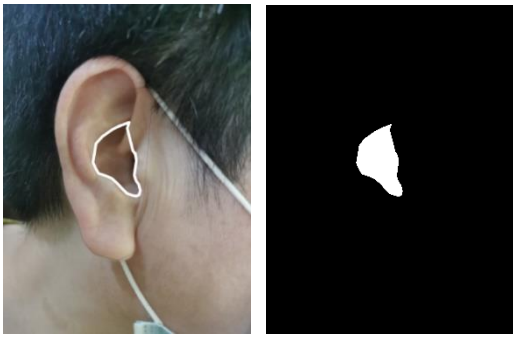
According to the latest Chinese national standards for ear acupoints GB/T13734-200891, there are 91 characteristic points for each ear image to achieve a positioning and segmentation in the ear image. Connecting the 91 sequential characteristic points according to certain rules. Regions of ears and ears' acupoint areas can be obtained by regularly connecting the characteristic points.

Connecting characteristic points from No. 1 to No. 24 as a closed loop, edges of entire ears are easily formed. Ear images are processed by boundary filling according to the edges, and processed images are seen as corresponding reference images. Image processing effect as shown in Fig. 5. In addition, the reference images of ear-armor part are processed in the same way. The effect as shown in Fig. 6.



(a) Whole ear edge (b) Reference image

Figure 5. Reference image processing for whole ear.



(a) Ear-armor edge (b) Reference image

Figure 6. Reference image processing for ear-armor.

IV. Experimental Results

In order to verify the performance of the proposed DeepEar, this paper valuate ear segmentation results from both qualitative analysis and quantitative analysis, this paper compare the proposed method with U-Net, PSPNet and a simple Convolution Neural Network (CNN). In this section, training details will be supplemented firstly and then the experimental results of ear segmentation will be analyzed. In addition, ear-armor segmentation results from DeepEar are also revealed.

A. Implementation Details

The training set and testing set obtained from section III are applied in the experiment. This paper resizes images to size 352×288 and trained the proposed model using ADAM, in addition, loss function is set as L1 loss between output tensor and reference image. The learning rate is initialed to $1e^{-3}$ decreased by $1e^{-6}$ every 100 iterations, results after 100 epoches are extracted. Pytorch as well as anaconda are joint with Nvidia GTX 3080 GPU for programming.

B. Segmentation Results Evaluation

Values of pixels in output tensor can be regarded as a binary classification problem, pixels in ear are seen as positive class, and pixels in background are seen as negative class. Thus, four situations will appear in the actual classification: if a pixel is in ear area and is predicted to be positive, it is a True Positive (TP), if a pixel is in ear area but predicted to be negative, it is a False Negative (FN), if a pixel is in background area but predicted to be positive, it is False Positive (FP), if a pixel is in background area and predicted to be negative, it is a True Negative (TN). Generally, the accuracy is expressed as Eq. (1).

$$Ac = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

However, the accuracy loses its significance in evaluation of classification results when the positive samples differ greatly from the number of negative samples. Indicators including precision (Pr), recall (Re) as shown in Eqs. (2)–(3) are introduced in the evaluation, which make experimental analysis more comprehensive.

$$Pr = \frac{TP}{TP + FP} \tag{2}$$

$$Re = \frac{TP}{TP + FN} \tag{3}$$

Evaluation the harmonic average of Pr as well as Re, a comprehensive indication called harmonic measure (F1) is obtained as shown in Eq. (4). It is obvious that F1 is persuasive for evaluating segmentation results.

$$F1 = \frac{2 \times Pr \times Re}{Rr + \mathrm{Re}} \tag{4}$$

Besides, negative pixels are also supposed to be predicted, specificity (TNR) is applied to evaluate negative pixels as shown in Eq. (5).

$$TNR = \frac{TN}{FP + TN} \tag{5}$$

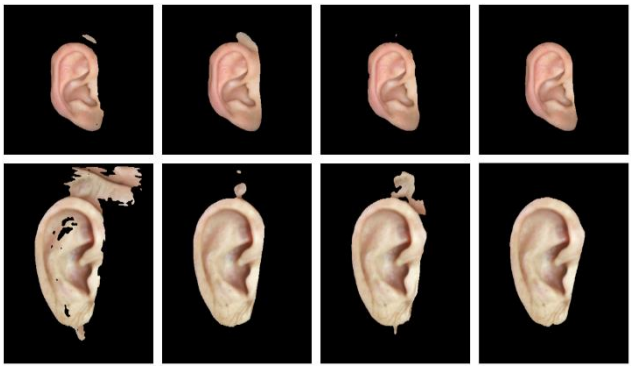
The testing set, including eight images from standardized images and 10 images from unstandardized images as well as their corresponding whole ear reference images, is used for results evaluation. Above five indicators including accuracy, precision, recall, F1 and specificity are applied in the quantitative analysis. The quantitative analysis results are presented in Table I.

TABLE I. Quantitative Analysis of Ear Segmentation Results

Method	Accuracy	Precision	Recall	F1	Specificity
CNN	0.9786	0.9333	0.9583	0.9430	0.9832
U-Net	0.9883	0.9549	0.9790	0.9650	0.9906
PSPNet	0.9868	0.9551	0.9721	0.9616	0.9896
DeepEar	0.9915	0.9762	0.9723	0.9738	0.9955

It is obvious that the proposed method obtains best results. Some of the segmentation images as shown in Fig. 7. The proposed DeepEar obtained the most complete segmented ear images without excess area.

This paper also made a segmentation in ear-armor area. Unfortunately, the segmentation results of other networks are convergence in a zero map. The ear-armor segmentation results from DeepEar as shown in Fig. 8. The proposed DeepEar basically completed ear-armor segmentation. However, some results exist excess areas not segmented, which is caused by the small target area and few training sets. As a result, there is still a lot of space for rise in the segmentation of ear images.



(a) CNN (b) U-Net (c) PSPNet (d) DeepEar

Figure 7. Ear segmentation results.

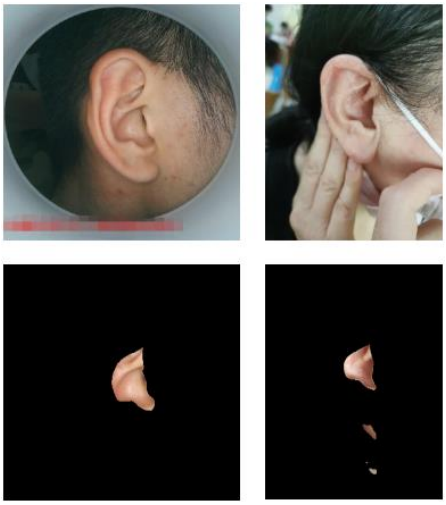


Figure 8. Ear-armor segmentation results from DeepEar.

V. Conclusion

In this paper, we proposed a DeepEar for ear segmentation. Combining spatial pyramid block and the encoder-decoder architecture, and retaining the atrous convolution from DeepLab, we made some adjustments for specific size requirements of ear images. Later on, we made experiments for ear segmentation. The results indicated that the proposed method achieved the best capability for ear segmentation. Meanwhile, we will hold on to collect ear datasets to improve the ear image segmentation effect. Moreover, the proposed method will be applied to acupoint diagnosis and treatment in intelligence Traditional Chinese Medicine to play an important role in healthcare engineering problems.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Yuhan Chen, Wende Ke, Yan Bai and Zhen Wang conducted the research; Qingfeng Li, Dongxin Lu provided the data; Yuhan Chen wrote the paper; all authors had approved the final version.

References

[1] Z. Fu, J. S. Dai *et al.*, "Analysis of unified error model and simulated parameters calibration for robotic machining based on lie theory," *Robotics and Computer-Integrated Manufacturing*, vol. 61, 101855, 2020.

[2] Y. Fu, H. Wang *et al.*, "A multifunctional robotic system toward moveable sensing and energy harvesting," *Nano Energy*, vol. 89, 106368, 2021.

[3] C. Hu, Q. Shi *et al.*, "Robotics in biomedical and healthcare engineering," *Journal of Healthcare Engineering*, pp. 1–2, 2017.

[4] Y. Hu, J. Li *et al.*, "Design and control of a highly redundant rigid-flexible coupling robot to assist the COVID-19 oropharyngeal-swab sampling," *IEEE Robotics and Automation Letters*, pp. 1856–1863, 2022.

[5] S.-Q. Li, W.-L. Guo, H. Liu *et al.*, "Clinical application of an intelligent oropharyngeal swab robot: implication for the COVID-19 pandemic," *Eur Respir J*, vol. 56, 2001912, 2020.

[6] S. Wang *et al.* "Design of a low-cost miniature robot to assist the COVID-19 nasopharyngeal swab sampling," *IEEE Transactions on Medical Robotics and Bionics*, vol. 3, no. 1, pp. 289–293, 2021.

[7] J. Tan, C. Fu *et al.*, "A flexible and fully autonomous breast ultrasound scanning system," *IEEE Transactions on Automation Science and Engineering*, pp. 1–14, 2022.

[8] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *International Journal of Computer Vision*, vol. 59, no. 2, pp. 167–181, 2004.

[9] Y. Huang and G.-Q. Yang, "Normalized cut segmentation algorithm combined with wavelet coefficient," *Journal of Computer Applications*, vol. 31, no. 1, pp. 182–242, 2011.

[10] M. Zhang, Y. Yang *et al.*, "Graph-cut based interactive image segmentation with randomized texton searching," *Computer Animation and Virtual Worlds*, vol. 27, no. 5, pp. 454–465, 2016.

[11] P. Joshi, R. K. S. Ranjth *et al.*, "Optic disc localization using interference map and localized segmentation using grab cut," *Automatika*, vol. 62, no. 2, pp. 187–196, 2021.

[12] N. Shah *et al.*, "k-Means image segmentation using mumford—Shah model," *Journal of Electronic Imaging*, vol. 30, no. 6, pp. 063029–063029, 2021.

[13] J. Li *et al.*, "Spectral clustering segmentation of high spatial resolution remote sensing imagery based on multi-scale object," *Science of Surveying and Mapping*, no. 10, 2019.

[14] J. Zhou *et al.*, "An adaptive meanshift segmentation method of remote sensing images based on multi-dimension features," *Geomatics and Information Science of Wuhan University*, vol. 37, no. 4, pp. 419–440, 2012.

[15] Y. Zhang *et al.*, "Semisupervised classification based on SLIC segmentation for hyperspectral image," in *Proc. IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 8, pp. 1440–1444, 2020.

[16] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[17] V. Badrinarayanan *et al.*, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[18] O. Ronneberger *et al.*, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015*, Springer International Publishing, 2015, pp. 234–241.

[19] H. Zhao, J. Shi *et al.*, "Pyramid scene parsing network," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[20] L.-C. Chen *et al.*, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," arXiv:1412.7062, 2016.

[21] L.-C. Chen *et al.*, "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, 2018.

[22] L.-C. Chen *et al.*, "Rethinking atrous convolution for semantic image segmentation," arXiv:1706.05587, 2017.

[23] L.-C. Chen *et al.*, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Computer Vision-ECCV 2018*, Springer International Publishing, 2018, pp. 833–851.

[24] W.-C. Hung *et al.* "Adversarial learning for semi-supervised semantic segmentation," arXiv:1802.07934, 2018.
[25] Y. Xue *et al.*, "SegAN: Adversarial network with multi-scale L1 loss for medical image segmentation," *Neuroinformatics (Totowa, N.J.)*, vol. 16, no. 3–4, pp. 383–392, 2018.
[26] E. Mussi *et al.*, "A novel ear elements segmentation algorithm on depth map images," *Computers in Biology and Medicine*, vol. 129, pp. 104157–104157, 2021.
[27] W. Ke, Y. Chen *et al.*, "An activate appearance model-based algorithm for ear characteristic points positioning," *Concurrency Computat. Pract. Exper.*, p. 7315, 2022.