A Survey on Deep Learning-based Eardrum Segmentation

Lu Zhang
Major in Bio Artificial Intelligence
Hanyang University
Ansan, Republic of Korea
kitsch@hanyang.ac.kr

Hoon Ji
ICT Department of Convergence Media Technology
Hanyang University
Ansan, Republic of Korea
greenpea0819@hanyang.ac.kr

Yeonjoon Lee
College of Computing
Hanyang University
Ansan, Republic of Korea
yeonjoonlee@hanyang.ac.kr

***Abstract*—Eardrum segmentation is an important part of the field of medical image segmentation, aiming to help doctors make accurate diagnoses and reduce their workload. Computer techniques have been widely used in this scenario. Due to the rapid development of deep learning, it has been replacing traditional computer vision methods and is widely used for segmenting medical images, several models have been used. In this paper, we investigate and analyze recent research related to deep learning on the eardrum segmentation scenario, introduce some basic background knowledge, and classify models, algorithms, and methods. We group the research by data feature. Finally, we speculate on the future direction of development and outlook.**

***Index Terms*—medical image segmentation, computer vision, deep learning**

I. Introduction

With the development of computer techniques, computer techniques have been widely used in the medical field. In the early years, deep learning [4] had not been widely used because of the limitation of the hardware, and because it was not complete, traditional image processing algorithms were used more. The advantage of the conventional computer vision method is stability, but it requires more experience, more experiments, and a higher workload. Over the past ten years, researchers and engineers have tended to use deep learning-related methods to solve such problems due to the widespread use of deep learning. Convolutional Neural Network(CNN) [1] has become the most popular used model because of its advantages, and there are also some other models used for such problems as Long Short Term Memory(LSTM) [2], Recurrent Neural Network(RNN) [3], etc.

Nowadays, benefits from the development of techniques, a huge amount of medical data can be stored and processed. In medical data, medical imaging is an essential part. The segmentation of the medical image is intended to replace traditional manual methods with human labor. Popular medical image segmentation tasks include various body parts, organs, and features, including eardrums. This survey will introduce some important models that have been implemented in eardrum segmentation, categorizing them by supervised type and backbones. For eardrum detection, current research is divided into two important parts mainly: in 2-dimension and 3-dimension. A 2-dimensional convolutional neural network cannot learn temporal information in the third dimension, and a 3-dimensional convolutional neural network often requires high computation costs and severe GPU memory resources. In this survey, we will focus on segmentation in 2-dimension.

II. Background

In the recent 20 years, benefits from the rapid development of computer techniques, many workloads that originally required manual processing have been replaced by computers for processing. As an important part of medical image segmentation, Eardrum segmentation has also been applied to numerous methods for decades ago. In the recent ten years, the rapid development of deep learning boosted it.

A. Traditional computer vision methods

Before the widespread implementation of deep learning, people were more inclined to use traditional image processing algorithms to detect eardrum boundaries. For example, in 2007, Liu et al. [21] used the canny algorithm to detect inner ear boundaries and the watershed algorithm to detect outer ear boundaries.

B. Deep Learning

Deep learning simulates the human neural network, abstracts the original data layer by layer by combining multiple non-linear processing layers, obtains different levels of abstract features from the data, and uses them for classification prediction. The advantage of deep learning is to use unsupervised or semi-supervised feature learning and hierarchical feature extraction efficient algorithms to replace manual feature acquisition.

1) Convolutional Neural Network(CNN): The basic structure of CNN consists of an input layer, a convolutional layer, a pooling layer (also called a sampling layer), a fully connected layer, and an output layer [5]. Generally, several convolutional layers and pooling layers are selected, and the convolutional layer and the pooling layer are alternately arranged; that is, a convolutional layer is connected to a pooling layer, and a convolutional layer is connected after the pooling layer, and so on. Since each neuron of the output feature map in the convolutional layer is locally connected to its input. The corresponding connection weight is weighted and summed with the local input. The bias value is added to obtain the neuron input value; the process is equivalent to the convolution process.

2) Backbone: Backbone is used for feature extraction and represents a part of the network. It is generally used to extract image information at the front end and generate a feature map for subsequent networks. VGGNet [6] and Resnet [7] are widely used because these backbone feature extraction capabilities are powerful, and official model parameters can be loaded and trained on large data sets (Pascal [22], Imagenet [23], etc.), and then connect to another network to fine-tune.

3) Transfer learning: Transfer learning tries to use unlabeled data in certain tasks or domains most efficiently. This is also the principle that semi-supervised learning abides by. Semi-supervised learning follows the settings of classic machine learning, but it only uses a limited amount of labeled data for training. In this way, semi-supervised domain adaptation essentially means semi-supervised learning under the condition of domain changes. Many lessons and ideas from semi-supervised learning apply equally to transfer learning.

4) Semantic segmentation: In medical image segmentation, the most representative methods are instance segmentation and semantic segmentation. For eardrum segmentation, semantic segmentation is the most used. Fig.1 [24] shows the difference between them. Semantic segmentation is the best method in tympanic membrane detection because it finds clear boundaries of the object. Also, because eardrum segmentation aims to find tympanic membrane only, instance segmentation is unsuitable in this scenario. Semantic Segmentation is a core field of computer vision, which classifies all pixels of a picture into a predetermined number of classes. It has developed rapidly with the help of deep learning.

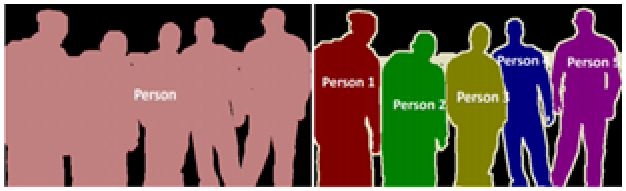

Fig. 1. The main difference between Semantic Segmentation and Instance Segmentation (The left image is Semantic Segmentation and the right one is Instance Segmentation) [24].

III. Eardrum Segmentation by Deep Learning

With the development of hardware, data, and algorithms, the practicability of deep learning has been enhanced. Compared to conventional computer vision methods, it usually takes less workload whereas performs well, so now, in most cases, techniques of deep learning segment medical images. By supervised types, it can be categorized into four categories, as shown in TABLE1.

TABLE I

An overview of different data features of different methods

Category	Data Feature
Supervised learning	Labeled
Semi-supervised learning	Labeled + Unlabeled
Weakly-supervised learning	Weakly labeled
Unsupervised learning	No label

A. Supervised learning

Supervised learning requires a large amount of carefully and accurately labeled data. Although it is difficult to obtain this kind of data, it is stable and has a high prediction rate and accuracy, so it is the most widely used method in eardrum segmentation.

FCN [1] evolved from CNN. The main difference between them is that FCN replaces the fully connected by the convolutional layer, and the final outputs are different, as shown in Fig.2. Regarding image segmentation algorithms, there are many traditions (about the history of image segmentation).

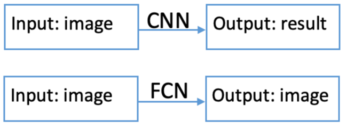

Fig. 2. Main difference between CNN and FCN

Pham et al. [10] proposed an approach for automatic segmentation, which is based on the FCN structure proposed. To train the network, they proposed a loss function combining AC loss with Dice loss. While trained by the proposed loss, the network can obtain better performances than using AC loss or Dice loss solely. It achieved good results in specific data sets.

FCN is the first time that the end-to-end convolutional network has been extended to the task of semantic segmentation. Many successful image segmentation deep learning technologies are based on the network structure implemented by FCN, such as U-Net [8], DenseNet [9], etc. These FCN-derived models can be regarded as independent individuals because they are too widely used. We will discuss them separately below:

1) U-Net: Researchers use encoder-decoder structure mostly like U-Net[8], DeepLab[11], etc. U-Net is the first high-impact encoder-decoder structure and the most used now, even became the baseline, it solves some disadvantages of general CNN, its structure is shown in Fig.3 [8].

It inspired many researchers to think of U-like semantic segmentation networks. Lots of research is related to modifications on U-Net.

Based on U-Net, Pham et al. [12] proposed EAR-UNet. It composes three main paradigms: EfficientNet for encoder, Attention gate for skip connection, and ResNet for the decoder. Compared to the original U-Net, they add attention module, integrate EfficientNet into the encoder and the Residual blocks into the decoder of the U-Net architecture, utilize the attention gate in the skip connection path to handle the variety in the shape of objects which are interested in, and introduce a new loss function based on shape distance for fully convolutional network training. They achieve better accuracy than other backbones in their dataset.

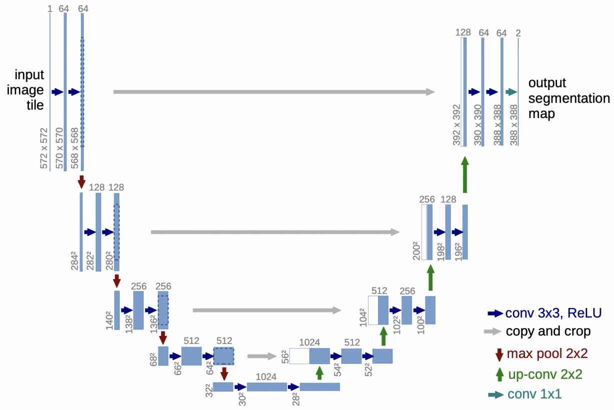

Fig. 3. U-net architecture (example for 32x32 pixels in the lowest resolution). Each bluebox corresponds to a multi-channel feature map. The number of channels is denotedon top of the box. The x-y-size is provided at the lower left edge of the box. Whiteboxes represent copied feature maps. The arrows denote the different operations [8].

2) DeepLab: Chen et al. [11] modified FCN and proposed this structure. It is widely used in medical segmentation. However, we did not find any eardrum segmentation research utilizing this structure.

3) DenseNet: The network structure based on CNN has become deeper. From AlexNet [14], VGG [6], GoogLeNet [15] to ResNet [7], researchers have never given up on the pursuit of depth, because deepening the network will indeed bring richer expressions, but as ResNet shows, as it turns deeper and deeper networks are not conducive to the propagation of gradients between layers. It is easy to cause vanishing gradient and exploding gradient. To solve this problem, DenseNet [9] uses a denser connection method which uses a forward propagation method to connect each layer with the rest densely. The purpose of this is to ensure that the information flow between the layers is maximized and all layers (feature map size matching) are directly connected. It reduces the vanishing gradient problem, enhances the spread of features between networks, realizes and strengthens feature reuse, and effectively reduces the number of parameters.

Khan et al. [16] implemented DenseNet to detect tympanic membrane and middle ear infections. They used several CNN architectures, fine-tuning them by PyTorch, replacing the softmax layer with three units. By comparing the results of different architectures, DenseNet, especially DenseNet161, performed the best in their dataset.

4) Recurrent Convolutional Neuron Network (RCNN): CNN is mainly for classification problems, whereas RCNN [17] is mainly for object detection. In RCNN, CNN is forced to focus on one area at a time because this minimizes interference. After all, only one object of interest is expected to dominate in a given area. The regions in RCNN are detected by a selective search algorithm and then resized to have the same size before they are fed into the CNN for classification and bounding box regression. Unlike other segmentations cited in the survey are semantic segmentation, RCNN is instance segmentation.

Seok et al. [18] pre-trained on MScoco [25] dataset then used Mask RCNN with ResNet50 as the backbone. It can improve accuracy either by sharpening the border or by improving the accuracy of the method of defining the shape of the malleus. However, they also found some limitations: the small size of the set of images, the data is images instead of videos (however, it can handle most scenarios of eardrum segmentation: video data is not common in this scenario), and parameter tuning was not done.

B. *Semi-supervised learning*

Unlike supervised learning, it is not widely used in eardrum segmentation because eardrum segmentation's result requires high accuracy, so although supervised learning requires labeled data, it is still the most popular method. However, in some scenarios(e.g., data that is hard to be labeled), it can also be used and perform better than supervised learning, such as the results of a hospital's examination. Doctors need a period of time to judge whether patients are healthy or not. There may be only a few sets of data to know if they are healthy or unhealthy; others only have data and don't know if they are healthy. So semi-supervised learning through the combination of supervised learning and unsupervised learning comes into researchers' views.

Cha et al. [19] applied transfer learning based on Inception-V3 [13] and ResNet101 for eardrum boundary detection. It pre-trained on ImageNet and selected two best-performed models out of 9 (The evaluation criteria are based on accuracy). This is the first study to utilize a deep learning scheme to classify tympanic membrane otoendoscopic images into six diagnostic categories.

C. *Weakly supervised learning and unsupervised learning*

Although weakly supervised learning requires less labeled data and unsupervised learning does not require labeled data, the difficulty of learning increased and the accuracy decreased, so currently, we have not seen any eardrum segmentation paper by weakly-supervised learning or unsupervised learning.

D. *Long Short Term Memory (LSTM)*

Most researchers use CNN-based models to process images, and LSTM is most used for natural language processing; however, we still found some papers using LSTM to segment eardrum boundaries. Ucar et al. [20] detected ear diseases by detecting RGB and key points by LSTM. The average accuracy was 99.06% for four different classes of the tympanic membrane. It follows the obtained keypoint positions, extracts hypercolumn deep features from 5 different layers of the VGG 16 model.

IV. DISCUSSION

We think the eardrum segmentation can be developed further: most research uses supervised learning. Although it performs well, it requires much carefully labeled data by manual labeling, which means a pretty much workload. In the future, there may be some new techniques that can replace the status of supervised learning.

Most studies have shown good accuracy in specific data. Still, there is no uniform horizontal comparison standard between studies due to different data sets, so we could barely say that a certain study is better than others based on accuracy. And with accuracy only, we cannot account for the performance of an algorithm because of potential overfitting problems. Therefore, a research result is only used on the data specified by its developer, and there is no result of extensive data application. This research field needs a more general study to identify a broader range of eardrum contour pictures to achieve performance improvements in practical applications.

V. CONCLUSION

In this survey, we can find the accuracy of deep learning is close to real doctors, or even better. In the future, it may take more workload and replace more workload from doctors and staff. We have discussed recent research for eardrum segmentation extended from deep learning, including the properties of networks, background knowledge, and others. The goal of this paper is to summarize and compare different methods to segment eardrum boundaries. We summarized why supervised learning has been widely used in this scenario and why U-Net is so popular. We can see most of the research has similarities: use CNN or modified CNN, merge different algorithms, add layers and modules like attention, merge different backbones, fine-tune softmax. Most research uses supervised learning. Unsupervised learning and weakly supervised learning still have room for development. We talked about U-Net, DeepLab, and its modified versions which the basic algorithm of segmentation is FCN. To satisfy more requirements in segmentation, DeepLab was invented by Google and applied to many scenarios. Meanwhile, in one case of the research, LSTM output an incredibly high accuracy. Overall, different techniques have different applicable scenarios, so how to choose the applicable technique according to the scenario is a question worth pondering.

REFERENCES

[1] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3431-3440.

[2] S. Hochreiter, J. Schmidhuber, Long short-term memory, Publisher, City, 1997.

[3] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning representations by back-propagating errors, Publisher, City, 1986.

[4] W.S. McCulloch, W. Pitts, A logical calculus of the ideas immanent in nervous activity, Publisher, City, 1943.

[5] R. Yamashita, M. Nishio, R.K.G. Do, K. Togashi, Convolutional neural networks: an overview and application in radiology, Publisher, City, 2018.

[6] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, Publisher, City, 2014.

[7] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, Publisher, City, 2015.

[8] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer, 2015, pp. 234-241.

[9] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700-4708.

[10] V.-T. Pham, T.-T. Tran, P.-C. Wang, M.-T. Lo, Tympanic membrane segmentation in otoscopic images based on fully convolutional network with active contour loss, Publisher, City, 2021.

[11] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Semantic image segmentation with deep convolutional nets and fully connected crfs, Publisher, City, 2014.

[12] V.-T. Pham, T.-T. Tran, P.-C. Wang, P.-Y. Chen, M.-T. Lo, EAR-UNet: A deep learning-based approach for segmentation of tympanic membranes from otoscopic images, Publisher, City, 2021.

[13] L.-C. Chen, G. Papandreou, F. Schroff, H. Adam, Rethinking atrous convolution for semantic image segmentation, Publisher, City, 2017.

[14] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Publisher, City, 2012.

[15] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1-9.

[16] M.A. Khan, S. Kwon, J. Choo, S.M. Hong, S.H. Kang, I.-H. Park, S.K. Kim, S.J. Hong, Automatic detection of tympanic membrane and middle ear infection from oto-endoscopic images via convolutional neural networks, Publisher, City, 2020.

[17] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2961-2969.

[18] J. Seok, J.-J. Song, J.-W. Koo, H.C. Kim, B.Y. Choi, The semantic segmentation approach for normal and pathologic tympanic membrane using deep learning, Publisher, City, 2019.

[19] D. Cha, C. Pae, S.-B. Seong, J.Y. Choi, H.-J. Park, Automated diagnosis of ear disease using ensemble deep learning with a big otoendoscopy image database, Publisher, City, 2019.

[20] M. Uçar, K. Akyol, Ü. Atila, E. Uçar, Classification of Different Tympanic Membrane Conditions Using Fused Deep Hypercolumn Features and Bidirectional LSTM, Publisher, City, 2021.

[21] https://cdmd.cnki.com.cn/Article/CDMD-10611-2009048155.htm

[22] http://host.robots.ox.ac.uk/pascal/VOC/

[23] https://image-net.org/

[24] Varatharasan, V., Shin, H. S., Tsourdos, A., Colosimo, N. (2019, November). Improving Learning Effectiveness For Object Detection and Classification in Cluttered Backgrounds. In 2019 Workshop on Research, Education and Development of Unmanned Aerial Systems (RED UAS) (pp. 78-85). IEEE.

[25] https://cocodataset.org/home