MULTI-MODAL AND MULTI-DIMENSIONAL BIOMEDICAL IMAGE DATA ANALYSIS USING DEEP LEARNING

A Dissertation

presented to

the Faculty of the Graduate School

at the University of Missouri

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

---

by

YANGYANG WANG

Dr. Filiz Bunyak Ersoy, Dissertation Supervisor

MAY 2023

The undersigned, appointed by the dean of the Graduate School, have examined the dissertation entitled

MULTI-MODAL AND MULTI-DIMENSIONAL BIOMEDICAL IMAGE DATA ANALYSIS USING DEEP LEARNING

presented by Yangyang Wang,
a candidate for the degree of Doctor of Philosophy,
and hereby certify that, in their opinion, it is worthy of acceptance.

Professor Filiz Bunyak Ersoy, chair

Professor Kannappan Palaniappan, member

Professor Teresa E. Lever, member

Professor Yunxin Zhao, member

To my parents, none of this would have been possible without your unconditional love and support, from my first day on earth to these last years spent so far away from you.

To Shizeng Yao, my doctor husband, who worked on his Ph.D. almost at the same time as I did and gives me great encouragement and accompaniment during this long and arduous journey.

ACKNOWLEDGMENTS

I want to express my gratitude to my advisor Dr. Filiz Bunyak for her advice, and support during all these years as a graduate student. I learned a lot as her student and feel particularly well prepared for my next endeavors thanks to her. I also want to thank Dr. Lever Teresa, who made it possible for me to work on such an interesting research topic.

I would like to thank Dr. Palaniappan, Dr. Zhao, and Dr. Lever for being members of my doctoral committee and for the invaluable insights they provided to improve this research work. I want to thank Dr. Palaniappan for all the particular help and support he provided while I worked with him at the CIVA lab. I want to thank Dr. Zhao for all the help and advice she provided while I worked as a teaching assistant for her course. I also want to thank Dr. Olga V Glinskii for offering me a great opportunity to work on the confocal microscopy project.

I would like to thank those people who gave me help during my Ph.D. study, thank Dr. Ali Hamad, Brianna McCarthy, Ashley Kloepper, Allison Jarombek, Erin White, and Elise Henn.

TABLE OF CONTENTS

ACKNOWLEDGMENTS . . . . . **ii**

LIST OF TABLES . . . . . **vii**

LIST OF FIGURES . . . . . **ix**

ABSTRACT . . . . . **xiv**

Chapter

1. INTRODUCTION **1**

1.1 Key Contributions 3
1.2 Thesis Outline 7

2. RELATED WORKS **9**

2.1 Oral-DDK syllable detction in oral-DDK test audio. 9
2.2 Laryngeal endoscopy video analysis 10
2.3 Dealing with objects in different scales 11
2.4 Orthogonality 12
2.5 Dealing with imbalanced data 12
2.6 Spatio-temporal networks for classification 13
2.7 Segmentation of meningeal microvasculature in confocal microscopy images 14
2.8 Class activate map 14
2.9 Weakly-supervised object localization and segmentation 15

3. EVENT DETECTION: DEEPDDK FOR ORAL DIADOCHOKINESIS ANALYSIS **16**

3.1 Introduction 16
3.2 DeepDDK Oral Diadochokinesis Analysis Network 19

3.2.1 Network Architecture 19
3.2.2 Convolutional Neural Networks Training of DeepDDK 20
3.2.3 Convolutional Neural Networks Testing of DeepDDK 22

3.3 Experimental Results of DeepDDK 24
3.4 MS-DeepDDK Multi-scale Oral Diadochokinesis Analysis Network 25

3.4.1 Visual Oral-DDK Analysis through Mouth and Jaw Motion Tracking 29

3.5 Experimental Results of MS-DeepDDK 30

3.5.1 Oral Syllable Detection Accuracy 30
3.5.2 K-folds Cross-validation of Syllable Event Detection of MS-DeepDDK 34
3.5.3 Visual Oral-DDK Analysis 36

3.6 Conclusion 36

4. VIDEO SEGMENTATION: ORTHOGONAL REGION SELECTION AND LARNET-STC NETWORKS 38

4.1 Introduction 38
4.2 Related Work 43

4.2.1 Laryngeal Endoscopy Video Analysis 43
4.2.2 Objects in Different Scales 45
4.2.3 Orthogonality 46
4.2.4 Imbalanced Data 46
4.2.5 Spatio-temporal Networks for Classification 46

4.3 Method 47

4.3.1 VFs State Estimation Network (LARNet) 48
4.3.2 Image Preprocessing and Subregion Generation 49
4.3.3 Orthogonal Region Selection Subnetwork (ORS) 49
4.3.4 Temporal Context-based Orthogonal Region Selection Network of LARNet-STC 51
4.3.5 Dealing with Imbalanced Data 53

4.4 Experimental Results of LARNet 53

4.4.1 Ablation Study of LARNet 54
4.4.2 Results of LARNet 56

4.5 Experimental Results of LARNet-STC 57

4.5.1 Ablation Study of LARNet-STC 58
4.5.2 Results of LARNet-STC 64

4.6 Conclusion 76

5. 3-D VOLUME SEGMENTATION: ENSEMBLE OF DEEP LEARNING CASCADES WITH GLOBAL SOFT ATTENTION 80

5.1 Introduction 80
5.2 Ensemble of Deep Learning Cascades for Vessel Segmentation 83

5.2.1 Image Preprocessing 83
5.2.2 Network Architecture 83
5.2.3 Decision Fusion (Late Classifier Fusion) 88

5.3 Experimental Results 88

5.3.1 Data Collection 88
5.3.2 Evaluation Metrics 89
5.3.3 Network Inference on 2-D Multi-focus Images 90
5.3.4 Network Inference on 3-D Image Stacks 93

5.4 Conclusion 94

6. DISCRETE FOURIER TRANSFORM CLASS ACTIVATION MAP (DFT-CAM) AND WEAKLY-SUPERVISED SEGMENTATION 97

6.1 Introduction 98
6.2 Related Works 100

6.2.1 Weakly-supervised Learning in Biomedical 100
6.2.2 Laryngeal Endoscopy Video Analysis 101

6.3 Method of DFT-CAM 102
6.4 Method of DFT-CAM Based Weakly-supervised Glottal Region Segmentation 105

6.4.1 Deep Classification Module 106
6.4.2 Weakly-supervised Glottal Region Segmentation 106

6.5 Experimental Results of Weakly-supervised Object Localization 107

6.5.1 Dataset 108
6.5.2 Weakly-supervised Object Localization (WSOL) 109
6.5.3 Evaluation of Weakly-supervised Object Localization 110
6.5.4 Ablation Study of DFT-CAM 112

6.6 Experimental Results of Weakly-supervised Glottal Region Segmentation 112

6.6.1 Dataset 112
6.6.2 Weakly-supervised Glottal Region Segmentation 113

6.6.3 Quantitative Evaluation of Segmentation 114
6.6.4 Qualitative Evaluation of Segmentation 118
6.6.5 Multi-task Analysis for Human Laryngeal Trans-nasal Endoscopy Video Data 118

6.7 Conclusion 120

7. CONCLUSION **122**

BIBLIOGRAPHY **125**

VITA **143**

LIST OF TABLES

Table **Page**

3.1 Layer details for CNN-1 and CNN-2 used for DDK syllable detection and localization. 21

3.2 DeepDDK's location accuracy of different types of syllables. 26

3.3 Average durations for oral-DDK syllables "Pa" & "Ta" & "Ka" and for various other syllables that are close substitutes. 28

3.4 Network parameters for the feature extraction streams in the proposed MS-DeepDDK network. 31

3.5 Mean absolute error (MAE) and mean square error (MSE) comparisons for oral-DDK syllable detection. 32

3.6 Cumulative scores (CS) for the compared methods for different count error tolerances. 32

3.7 Detailed cumulative score (CS) analysis for the proposed MS-DeepDDK network using five-folds cross-validation. 33

4.1 Distribution of the various challenging cases in the dataset. 41

4.2 VF state estimation performance. 55

4.3 Contribution of the different network components to the classification performance. 58

4.4 Feature summary of the proposed network and the comparison networks. The architectural differences between the four spatio-temporal context-based networks are illustrated in Figure 4.6. 59

4.5 Networks Features 59

4.6 VFs state estimation performance for the proposed and comparison networks. Features of the compared networks are given in Table 4.4. 62

4.7 Five-fold cross-validation results for the proposed LARNet-STC network. 63

4.8 The five-fold cross-validation results comparison of LARNet, LARNet-ST2Le-Conv, and the proposed LARNet-STC. The first value is mean, the second value is standard deviation. 67

4.9 Statistical significance analysis. The five-fold cross-validated paired t test of the F1 scores of three different networks. 68

4.10 Segmentation-derived LAR/non-LAR classification results. Average F1 scores for non-LAR frames. 72

4.11 Segmentation-derived LAR/non-LAR classification results. Average F1 scores for LAR frames. 72

5.1 Number of trainable parameters of the compared networks. 90

5.2 Single network segmentation performances. 95

5.3 Ensemble network segmentation performances. 96

6.1 Average Intersection over Union (IoU) (%) scores between the ground truth and the CAM-generated bounding boxes. Bold fonts mark the best results, underlined fonts mark the second-best results. 109

6.2 DFT-CAM weakly-supervised object localization IoU scores (%) for different k values. 111

6.3 Average Intersection over Union (IoU) (%) scores between the ground truth and the predicted bounding boxes of the glottal region. Bold fonts mark the best results. 112

6.4 Average Intersection over Union (IoU) (%) scores between the ground truth and the predicted glottal region masks. Bold fonts mark the best results. 115

6.5 Average Intersection over Union (IoU) (%) scores comparison between multiple glottal region segmentation methods 118

LIST OF FIGURES

Figure **Page**

3.1 Audio waveform samples for different types of oral-DDK tasks. 18

3.2 CNN-1 and CNN-2 architectures used for DDK syllable detection and localization. 20

3.3 Sample training data. Colored dots mark ground-truth timestamps, shaded regions mark positive training samples for CNN-1. 22

3.4 Intermediate outputs from the different stages of DeepDDK for a sample "Pa" file. Top panel: original audio signal (blue) with ground-truth timestamps (red). Second panel: output of CNN-1. Third panel: output of CNN-2 where local maxima indicate syllable timestamp. 23

3.5 Cumulative distribution of event count error for pre-linguistic segmentation[1], Smekal et al.[2][3], MFCC with Linear SVM and our DeepDDK software. Horizontal axis indicates count error (difference between the number of predicted events vs. ground truth events). Vertical axis shows the ratio of the test files. Absolute event count differences of 1, 2, 3, 4, 5 in the graph correspond to percent count errors of 1.35%, 2.70%, 4.05%, 5.40%, 6.75%, respectively (average number of events per file is 74). 25

3.6 The proposed multi-modal and multi-scale oral-DDK analysis pipeline. 26

3.7 The architecture of the proposed multi-scale syllable detection deep learning network MS-DeepDDK. 27

3.8 Sample screenshot for JawTrack visual tracking software during oral-DDK mouth and jaw motion analysis. 30

3.9 Cumulative scores (CS) for different count error tolerances. Comparison between methods: pre-linguistic segmentation [1], Smekal et al. [2, 3], MFCC with Linear SVM, DeepDDK [4], and the proposed multi-scale DeepDDK. Absolute event count differences of 0, 1, 2, 3, 4, 5 in the x-axis of the graph correspond to percent count errors of 0%, 1.35%, 2.70%, 4.05%, 5.41%, 6.76%, respectively (using the average number of events per file as 74). 33

3.10 Boxplot of the absolute count errors for five-folds cross-validation. 34

3.11 Audio-visual analysis of oral-DDK tests. Top: The blue waveform is the original audio waveform. The green line on top of the blue waveform is the signal envelope. Yellow dots mark event timestamps generated by the proposed MS-DeepDDK network. Bottom: the blue signal denotes the distance between the philtrum and mentolabial sulcus in time, automatically computed using our JawTrack visual tracking software. Green and red dots represent opened and closed states of the mouth/jaw outputted by JawTrack. 35

4.1 Sample images for the three vocal fold state classes. 39

4.2 Sample laryngoscopy video frames illustrating different processing challenges. (a-c) Left images show original video frames, right images show corresponding histogram equalized images. 43

4.3 Architecture of the proposed VFs state estimation network. 48

4.4 Subregion cropping and the Orthogonal Region Selection (ORS) subnetwork. Inputs to the network are five cropped subregions (marked with yellow squares) from the preprocessed image. Output of the network is a 1-D feature vector corresponding to the selected subregion. This vector is selected from F by the index j^* of the minimum value in O. “FC” represents fully-connected layer. 50

4.5 Architecture of the proposed spatio-temporal context-based orthogonal region selection network. On top of the VF state estimation networks, a set of fully convolutional layers are inserted to the network to incorporate temporal context. “Conv” represents convolution operation. 51

4.6 Four different architectures of spatio-temporal context-based networks. 61

4.7 Boxplot of the F1 scores for the five-fold cross-validation of three proposed networks. The green triangle is the mean across five folds. 69

4.8 Quantification evaluation of LAR event durations (number of frames). (a) Histogram of the distribution of ground truth LAR event durations. (b) cumulative distribution of the frame error of LAR event prediction. (c) Comparison of the ground truth and prediction of VF states for a single video. (d) Sample original video frames at timestamps A, B, C, and D in (c). 69

4.9 Segmentation-derived LAR/non-LAR classification results. Average F1 scores for non-LAR frames. VFs segmentation algorithms (U-LSTM [5], FCRN [6], and FCRN [6] + histogram equalization + ORS) and the proposed LARNet-STC. 73

4.10 Segmentation-derived LAR/non-LAR classification results. Average F1 scores for LAR frames. VFs segmentation algorithms (U-LSTM [5], FCRN [6], and FCRN [6] + histogram equalization + ORS) and the proposed LARNet-STC. 73

4.11 Visual explanation of the LARNet-STC network output using Grad-CAM visualization [7]. Top row: subregions automatically selected by Orthogonal Region Selection (ORS) subnetwork. Bottom row: regions corresponding to high score for the predicted class marked with highlights changing from red to blue corresponding to higher to lower impact regions. 74

4.12 The confusion matrix of the results from the proposed context-based orthogonal region selection network (LARNet-STC). 76

4.13 Sample outputs from the proposed system. Red label represents ground truth. Green label represents prediction. 77

4.14 Sampled non-LAR sequential video frames (frame 6-10) with visual occlusion from laryngoscopy videos. 77

5.1 Sample blood microvascular structures imaged using confocal microscopy. The first three columns show sample single focus slices. The last column shows fused multi-focus image. 82

5.2 Sample confocal microscopy fused multi-focus images before (first column) and after (second column) adaptive histogram equalization. 84

5.3 Architecture of the proposed deep binary attention cascade (DBAC). 85

5.4 Architecture of the proposed deep distance map attention cascade (DDMAC). 85

5.5 Intermediate and final segmentation results for two sample multi-focus input images. (a) input images, (b-c) ground truth maps for the binary attention module in DBAC and dilated probability module in DDMAC, (d-f) predicted segmentation masks versus ground truth for DBAC, DDMAC, and ensemble networks. The red, blue, and white regions represent false-positive, false-negative, and true positive predictions respectively. 91

5.6 Sample outputs from the proposed system. (a) multi-focus confocal microscopy image enhanced with adaptive histogram equalization; (b) predicted segmentation mask versus ground truth where the red, blue, and white regions represent false-positive, false-negative, and true positive predictions respectively. (c-d) 3-D segmentation masks obtained by applying the proposed 2-D segmentation network to the individual single focus images forming the confocal microscopy volume. Visualization of the 3-D segmentation results were generated using the Chimera software [8]. Color fades with increasing depth. 92

6.1 Processing steps of the proposed discrete Fourier transform driven class activation map DFT-CAM. The figure illustrates selection of k representative channels out of m original channels through DFT-based feature encoding and orthogonality-based feature selection. Selected feature channels are then aggregated to produce the proposed class activation map. 102

6.2 Weakly-supervised glottal region segmentation. 104

6.3 Processing steps of the proposed Discrete Fourier Transform driven class activation map DFT-CAM. The figure illustrates selection of k representative channels out of m original channels through DFT-based feature encoding and orthogonality-based feature selection. Selected feature channels are then aggregated to produce the proposed class activation map. 105

6.4 UNet architecture with two prediction outputs, one is a binary mask for the glottal region, and the other one is a binary mask for the edge of the glottal region. Numbers above the convolutional blocks are the corresponding numbers of convolutional channels. 108

6.5 Sample weakly-supervised object localization results obtained from the VGG-16 classification network [9] outputs using the different class activation map (CAM) methods. The heatmaps overlayed on the original images illustrate the CAM results. The red to blue colormap represents high to low probabilities. All bounding boxes were generated from the class activation maps as described in Section 6.5.2. Each column shows the outputs from the same CAM method. Each row shows the same input image and its class label. The green and red bounding boxes correspond to the ground truth and CAM bounding boxes respectively. The number below each image represents the IoU score between the ground truth and CAM boxes. 111

6.6 Visualization comparison of UNet-DFT and UNet-Grad predictions. The yellow transparent mask is the UNet prediction. The solid grey mask is the glottal region segmentation ground truth mask. The index below each image indicates "#video sequence index - frame index: class label" of that image. (a) contains segmentation outputs predicted by the two-outputs UNet that is trained using DFT-CAM's outputs. (b) contains segmentation outputs predicted by the two-outputs UNet that is trained using Grad-CAM's outputs. Some of the frames don't have glottal region segmentation ground truth masks because they are either LAR (VFs closed) or occluded, and the vocal fold is not visible in these cases. 116

6.7 Output of the multi-task laryngeal analysis system for a low-speed transnasal endoscopy video. The X-axis is the video frame index. The Y-axis is the glottal region size. The blue signal represents the classification result-corrected glottal region sizes. The orange signal represents ground truth glottal region sizes. The green dash represents the UNet segmented glottal region sizes. The red signal represents the ground truth classification label, where the low signal is "non-LAR" class, the median-high signal is "LAR" class, and the highest signal is "occlusion" class (note that the ground truth classification label doesn't reflect the glottal region size of the corresponding frame). The grey region represents the predicted "LAR" class. The green region represents the predicted "occlusion" class. The white region besides grey and green represents the predicted "non-LAR" class. 119

MULTI-MODAL AND MULTI-DIMENSIONAL BIOMEDICAL IMAGE DATA ANALYSIS USING DEEP LEARNING

Yangyang Wang

Dr. Filiz Bunyak Ersoy, Dissertation Supervisor

ABSTRACT

There is a growing need for the development of computational methods and tools for automated, objective, and quantitative analysis of biomedical signal and image data to facilitate disease and treatment monitoring, early diagnosis, and scientific discovery. Recent advances in artificial intelligence and machine learning, particularly in deep learning, have revolutionized computer vision and image analysis for many application areas. While processing of non-biomedical signal, image, and video data using deep learning methods has been very successful, high-stakes biomedical applications present unique challenges such as different image modalities, limited training data, need for explainability and interpretability etc. that need to be addressed.

In this dissertation, we developed novel, explainable, and attention-based deep learning frameworks for objective, automated, and quantitative analysis of biomedical signal, image, and video data. The proposed solutions involve multi-scale signal analysis for oral-diadochokinesis studies; ensemble of deep learning cascades using global soft attention mechanisms for segmentation of meningeal vascular networks in confocal microscopy; spatial attention and spatio-temporal data fusion for detection of rare and short-term video events in laryngeal endoscopy videos; and a novel discrete Fourier transform driven class activation map for explainable-AI and weakly-supervised object localization and segmentation for detailed vocal fold motion analysis using laryngeal endoscopy videos. Experiments conducted on the proposed methods showed robust and promising results towards automated, objective, and quantitative analysis of biomedical data, that is of great value for potential early diagnosis and effective disease progress or treatment monitoring.

CHAPTER 1

INTRODUCTION

Biomedical images that create visual and functional representations of the interior of the human body provide indispensable non-invasive diagnostic capabilities in the modern healthcare system [10]. Thanks to the advances and increased availability of the medical imaging devices, biomedical image data from sub-cellular to organ scales, involving different modalities (i.e. electron and light microscopy, endoscopy, computed tomography, magnetic resonance imaging etc.) is growing at an unprecedented rate. Novel image analysis methods and tools are needed to take full advantage of this data. Automated, quantitative, and objective analysis of biomedical image data can aid in early diagnosis, disease progress monitoring, and treatment efficacy monitoring, which are crucial for medical decision making, timely intervention, and ultimately, improving patients' quality and duration of life. However, biomedical image analysis remains to be challenging, due to (1) high-stakes nature of the application that requires higher levels of accuracy, explainability and interpretability; (2) unique image modalities; (3) complex anatomical structures; (4) large variability in the data; and (5) very limited training data due to expertise requirements, labor-intensive nature of the process, and lack of intuitive annotation tools for complex 2-D + time or 3-D volumetric data.

In recent years, deep learning, a subfield of artificial intelligence and machine learning (AI/ML), has emerged as a powerful tool for image processing and computer vision. Deep learning architectures are composed of multi-layer stacks of simple modules, all (or most) of which are subject to learning, and many of which compute non-linear input-output

mappings [11]. Deep learning has substantially improved state-of-the-art in many image processing and computer vision tasks for both general and biomedical applications [12] [10] [13] [14] [15]. However, two issues, the black-box nature of classical deep learning methods and their need for large amounts of training data, have restricted their use in clinical applications. Very recently, the Attention Model (AM), first introduced for machine translation [16] [17], has become increasingly popular within the artificial intelligence (AI) community. Attention model serves as a resource allocation scheme and is becoming an essential component of neural architectures especially deep learning architectures for a remarkably large number of applications including natural language processing (NLP), speech recognition, and computer vision (CV) [18] [16]. Beyond improving performance of neural networks, recently, attention mechanism has also been used as a tool to help improve interpretability as well as transparency and fairness of neural network architectures.

The field of explainable artificial intelligence (XAI) focuses on understanding and interpretation of AI systems. XAI aims to address the need for trustworthy, fair, robust, high performing models for real-world applications [19] [20]. XAI methods attempt to (1) explain the decisions made by algorithms; (2) unravel the patterns within the inner mechanism of an algorithm; and/or (3) present the system with coherent models or mathematics [21]. By applying different XAI methods such as visualization of artificial intelligence behaviors, developing white-box models, building logical flow graph explanations etc., artificial intelligence systems can be made more interpretable and reliable for humans. Such systems can better serve the decision-making process in the biomedical field.

For optimum performance, deep learning systems require large amounts of annotated data. In biomedical fields, lack of annotated training data tends to be more severe compared to the other application areas, because of difficulties in data acquisition and requirement of expertise for data labeling. Weakly-supervised learning, that aims to reduce data annotation workload while still keeping comparable accuracy and precision levels in the deep learning network output, started to receive enormous attention. Recently, the class activation map

(CAM) technique, which was originally developed for XAI, has been proposed to implement weakly-supervised learning. The CAM methods generate a discriminate saliency map for a specific class from the deep classification networks, showing the pixel-wise probability of a pixel being used for the final class label prediction. The larger the probability, the more likely it is that the pixel belongs to the target object. This transformation from an image-level class label to a pixel-level class activation map enables pixel-wise precision outputs by only using image-level inputs and leads to considerable savings in data annotation workload. Because of this possibility of reduction in annotation workload, CAM techniques started to be applied to weakly-supervised learning tasks such as weakly-supervised object localization and weakly-supervised object segmentation.

In this dissertation, we focused on developing novel, explainable, attention-based, weakly-supervised deep learning solutions for biomedical signal, image, and video analysis. We proposed solutions to three different types of data modalities and associated biomedical signal, images, and video analysis problems: (1) 1-D signal data analysis for speech and swallow studies involving oromotor system; (2) 2-D image and 3-D image volume analysis for meningeal microvascular system study; (3) 2-D + time (video) analysis for vocal folds motion study. We developed deep learning network architectures involving novel attention, information selection, and fusion mechanisms, and a novel frequency domain class activation map method. The proposed systems aims to enable automated, objective, and quantitative analysis of these three different data modalities, addressing specific needs of biomedical data analysis.

1.1 Key Contributions

In this dissertation, we made contributions to deep learning architectures, attention mechanisms, explainable AI (XAI), weakly-supervised object localization and segmentation, and information selection and fusion. We developed fully-supervised and weakly-supervised machine-learning-based algorithms that are able to handle 1-D signal, 2-D im-

age, 3-D volume, and 2-D + time (video) data. Through the developed methods, we facilitated automated, objective, and quantitative analysis of the oral diadochokinesis (oral-DDK) test, flexible endoscopic evaluation of swallowing studies, and meningeal microvasculature system. The developed methods also allowed extraction of quantitative biomarkers to facilitate scientific understanding, early diagnosis, monitoring of diseases and treatments.

The key contributions of this dissertation can be summarized as follows:

1. **DeepDDK: A Deep Learning based Oral-Diadochokinesis Analysis Software.** We proposed DeepDDK, a cascaded end-to-end trainable network for fast-speech audio event detection and timestamp prediction. The input to the DeepDDK is a 1-D audio signal acquired during oral-DDK tests. In order to label oral-DDK data, we have developed a preliminary unsupervised automated syllable detector (relying on the signal envelope and local maxima detection) with a user interface for visualization, navigation, and modification of the results. Annotation consists of a single event timestamp for each audio event and doesn't include any information regarding associated temporal interval (i.e. event start and end timestamps). The first subnetwork of DeepDDK predicts start and end timestamps for each event. The second subnetwork of DeepDDK predicts a single timestamp within each event.

2. **MS-DeepDDK: Multi-Modal and Multi-Scale Oral-Diadochokinesis Analysis using Deep Learning.** We proposed a multi-modal oral-DDK test analysis system involving automated processing of complementary 1-D audio and 2-D video signals acquired during speech and swallowing studies. The system aims to automatically generate objective and quantitative measures from the oral-DDK tests to aid early diagnosis and treatment monitoring of neurological disorders. The audio signal analysis component of the proposed system involves a novel multi-scale deep learning network. The video signal analysis component involves tracking mouth and jaw mo-

tions during speech tests using our visual landmark tracking software. This system is an extension and improvement of our original DeepDDK system.

3. **LARNet: Orthogonal Region Selection Network for Laryngeal Closure Detection in Laryngoscopy Videos.** We proposed a deep learning-based image analysis solution for automated detection of laryngeal adductor reflex (LAR) events in laryngeal endoscopy videos. The proposed system consists of a two-stream network with a novel orthogonal region selection subnetwork (ORS). The proposed network classifies each video frame into one of three vocal fold (VF) states: non-LAR (open VFs), LAR (closed VFs), and visual occlusion (the VFs are either masked/covered by other anatomical structures or out of the camera field of view). The proposed approach combines global and local information through a two-stream network and a novel orthogonal region selection (ORS) subnetwork that works like an unsupervised attention mechanism to improve VF state estimation accuracy.

4. **LARNet-STC: Spatio-Temporal Context-based Orthogonal Region Selection Network for Laryngeal Closure Detection in Laryngoscopy Videos.** We proposed LARNet-STC, a deep learning model that extends our novel orthogonal region selection network with temporal contextual information. This network learns to directly map its input to a VF open/close state without first segmenting or tracking the VFs , which drastically reduces labor-intensive manual annotation needed to generate segmentation mask or VF motion tracks. The proposed spatio-temporal context-based orthogonal region selection network allows the integration of local image features, global image features, and information on VF states in time for robust LAR event detection. This deep learning network is a further improvement of our original LARNet system.

5. **Ensemble of Deep Learning Cascades for Segmentation of Blood Vessels in Confocal Microscopy Images.** We proposed a deep learning system for robust segmen-

tation of cranial vasculature of mice in confocal microscopy images. The proposed system is an ensemble of two deep-learning cascades consisting of two coarse-to-fine subnetworks with skip connections in between. One cascade aims to improve sensitivity, while the other aims to improve the precision of the segmentation results. The proposed cascades first learn to predict two soft attention maps, one based on binary pixel classification, and the other based on regression to a distance map. Then, the attention maps guide the networks to predict an accurate vessel segmentation mask. To compensate for limited confocal microscopy training data, each of the proposed cascades is first trained with an epifluorescence microscopy image dataset [22][23], then fine-tuned with a small set of fused confocal microscopy images of mice cranial microvasculature.

6. **DFT-CAM: Discrete Fourier Transform Driven Class Activation Map.** We proposed a gradient-free, discrete Fourier transform driven, class activation map method named DFT-CAM. Discrete Fourier Transform (DFT) converts spatial domain information in the images into frequency domain. Frequency domain representation allows better separation of significant semantic information from image details and noise. These representations can be used to summarize geometrical characteristics of spatial information [24]. The proposed DFT-CAM method first uses discrete Fourier transform (DFT) based representation to summarize learned features in convolutional feature maps; then uses feature orthogonality to automatically select the most representative semantic features while preventing the inclusion of less-contributed features.

7. **Weakly-supervised Object Localization.** We proposed a DFT-CAM based weakly-supervised object localization pipeline, which uses image-level training labels to predict pixel-level detection outputs for target objects.

8. **Weakly-supervised Object Segmentation.** We proposed a weakly-supervised glottal region segmentation system, which doesn't require manually annotated segmentation ground truth for deep segmentation training. The proposed system involves three modules, a deep classification module, a pseudo binary mask generation module, and a weakly-supervised deep segmentation module. The deep classification module directly uses our previously proposed human laryngeal closure detection network, called LARNet-STC [25]. Then, our gradient-free, discrete Fourier Transform driven class activation map method named DFT-CAM is adopted to produce a class activation map (CAM) during the inference process of the LARNet-STC. Active contours [24] are used to generate pseudo binary masks of the glottal regions using the thresholded CAM results. Finally, the pseudo binary masks of glottal region are used to train a modified U-Net [26] with two complementary outputs to predict a segmentation mask of the glottal region.

1.2 Thesis Outline

The rest of the dissertation is organized as follows: Chapter 2 reviews the related work. Chapter 3 introduces 1-D signal event segmentation networks DeepDDK and Multi-scale multi-modal DeepDDK (MS-DeepDDK) that are trained with weak supervision. These two deep learning networks address the issue of lacking information about the start and end timestamps for each signal event. The proposed DeepDDK and MS-DeepDDK are designed to have light deep-learning architectures with efficient supervision strategies, aiming at fast and accurate training and inference processes. In this chapter, we show the evaluation results of DeepDDK and MS-DeepDDK on oral-DDK data.

Chapter 4 presents spatial-temporal orthogonal region selection (ORS) networks LARNet and LARNet-STC for video segmentation. First, we proposed an attention mechanism orthogonal region selection (ORS) subnetwork. Then, we proposed ORS-based LARNet and LARNet-STC. These two deep classification networks address the issue of imbalanced

classes of training data. In this chapter, we evaluate the LARNet and LARNet-STC on human the laryngeal transnasal endoscopy video dataset.

Chapter 5 presents an ensemble of deep learning cascades designed for 2-D image-based segmentation. The proposed segmentation system with two deep-learning cascades overcomes the issue of segmentation ground truth lacking and improves the efficiency of network training. The proposed pipeline is applied to confocal microscopy images of the meningeal microvascular system for evaluation.

Chapter 6 presents Discrete Fourier Transform driven class activation map (DFT-CAM) and DFT-CAM based weakly-supervised object localization and segmentation. First, we describe the Discrete Fourier Transform driven class activation map (DFT-CAM). Then, we describe the weakly-supervised object localization and segmentation pipelines based on the DFT-CAM. The weakly-supervised object localization is evaluated on the ImageNet Large Scale Visual Recognition Challenge 2012 (ILSVRC2012) dataset [27]. The weakly-supervised object segmentation is applied to the human laryngeal transnasal endoscopy video data for evaluation. In the end, together with the previously proposed LARNet and LARNet-STC, we describe the joint analysis of classification and segmentation results further improve the robustness and comprehensiveness of the visual-based analysis for the vocal fold system.

Chapter 7 concludes this dissertation and discusses future works. Methods presented in this dissertation, originating from collaborations with researchers in biomedical and computational sciences, have been published in [4] [28] [29] [30] [25].

CHAPTER 2

RELATED WORKS

2.1 Oral-DDK syllable detction in oral-DDK test audio.

Recent studies have started to show the potential of speech in general and oral diadochokinesis (oral-DDK) in particular to be a functional biomarker for neurological disorders. DDK task derived measures were explored for diagnosis of neurological diseases such as Parkinson's disease (PD) [31][32], traumatic brain injury [33], multiple sclerosis (MS) [34], and ataxic dysarthria [35]. Various computational approaches have been proposed to analyze DDK data and to derive clinically relevant outcome measures. For example, syllables in oral-DDK task can be detected by first computing signal envelope, then by thresholding the envelope or locating local maxima in the envelope. This process requires parameter selection for envelope computation and thresholding. However, complexity of the signals, and high variations in frequency and amplitude make parameter selection challenging and result in under- or over-detection of the syllables. Wang et al. [36] proposed a multi-threshold syllable detection system in which a threshold is automatically selected based on a 7-second DDK sample and the gender of the participant. Threshold can then be adjusted to re-perform the analysis if needed. However, if the lowest peak intensity during consonant-vowel (CV) is lower than the highest peak intensity during inter-syllable pauses, the DDK sample gets labeled as nonexecutable. The approach results in more than one third of their DDK samples being unanalyzable. Tao et al. [37] proposed use of Gaussian Mixture Models and Hidden Markov Models (GMM-HMM) to automatically detect syllable boundaries in DDK data. While there is significant interest and substantial success in

deep learning-based audio analysis systems [38][39][40], oral-DDK analysis remains to be a challenging task due to fast and repetitive nature of the data, need for accuracy, and lack of semantic context clues that can help the process.

2.2 Laryngeal endoscopy video analysis

Majority of the laryngeal endoscopy video analysis studies focus on processing of rigid transoral (through the mouth) laryngoscopy videos. Transoral laryngoscopy videos are characterized by high frame rates, high image resolutions, and good image quality. On the other hand, flexible transnasal (through the nose) endoscopy, used in clinical practice and focus of this dissertation, suffer from lower frame rates, image resolutions, and quality. The earlier works on laryngeal endoscopy video analysis mostly relied on sensitive handcrafted/engineered features that require heavy parameter tuning. Some examples of these works include [41][42][43] designed for vocal fold (VF) segmentation, [44] designed for VF classification , and [45] designed for VF image selection. Recently, deep learning-based approaches started to emerge for automated analysis of laryngeal videos. In [6], we have developed a deep convolutional regression network for segmentation of VFs and glottal region. Annotated training data for this network was generated by extending our previous interactive vocal fold tracking software VFTrack [46]. In [47], a cascade of two networks was proposed to segment the laryngeal structures. In [5], a U-Net [26] based segmentation network was augmented with Long Short-Term Memory (LSTM) [48] blocks for segmentation of glottal region from transoral laryngeal endoscopy videos. This dissertation focuses on detection of laryngeal adductor reflex (LAR) events in flexible, transnasal endoscopy videos. While great improvement over the previous handcrafted/engineered works, VF segmentation networks are not very suitable for LAR event detection for four main reasons: (1) segmentation networks require ground truth segmentation masks for training. Preparation of these training segmentation masks is a labor intensive and time-consuming task requiring domain expertise. Whereas frame level class labels (i.e. open VFs, closed VFs)

required for VF state description require much less effort and expertise. (2) Segmentation networks are often tested on video frames where VFs are fully visible. They often detect spurious regions when VFs are occluded or are not in the field of view of the endoscope. (3) Abrupt motion during LAR events leads to tracking failures. (4) Additionally, most of the earlier work on VF analysis did not incorporate temporal video context and analyzed individual video frames independently leading to loss of valuable temporal information for LAR event detection.

2.3 Dealing with objects in different scales

In deep convolutional neural networks, max-pooling layers designed to enlarge the convolution field of view play a critical role. They present a compromise between the hardware capability, execution time, and the convolutional kernel size. They help a network extract and learn abstract and transformation invariant features from an image. Unfortunately, small objects can also be eliminated in this process. If small objects are of interest to the given task, network performance can get adversely affected. In our applications, VFs typically occupy a small portion (typically $< 25\%$) of the endoscope field of view. In some cases, when the endoscope is positioned further away from the VFs, or when the endoscope is being pulled out of the larynx, the apparent size of the VFs can be even smaller. Furthermore, the relatively far distance between the VFs and the endoscope makes the camera harder to autofocus. Depending on the position and orientation of the endoscope, VFs can appear at different positions in a video frame. VF state estimation performance is expected to be adversely affected by the small size of VFs in these inspected images.

To deal with small object classification/segmentation problems, regions of interest are usually extracted from the input images before performing classification or segmentation tasks. For example, in deep object detection networks such as YOLOv4 [49] and feature pyramid network [50], first a set of bounding boxes are proposed, then image content within the proposed bounding boxes are classified and/or segmented. These detection net-

works require ground truth bounding boxes and segmentation masks besides class labels for training, which increase manual annotation workload and slow down the speed of the process. In this work, we propose an unsupervised region selection scheme (orthogonal region selection (ORS) subnetwork) that selects an image subregion without need for manual annotation.

2.4 Orthogonality

Orthogonality has been used for network initialization in deep neural networks (DNNs) [51][52]. Orthogonal weight normalization has been proposed to improve network generalization capabilities [53]. There is a growing interest on use of deep neural network layer weight pruning techniques to reduce feature redundancy in DNNs. The aim is to improve the generalization and precision capabilities of DNNs [54]. Orthogonality has been used in network initialization and regularization to prevent gradient vanishing or exploding problem in training very deep neural networks [55][56]. [57] proposed a loss function that encourages the features of different classes to be orthogonal to each other. Orthogonal deep features decomposition has been proposed to improve face recognition accuracy [58]. In this work, we relied on feature orthogonality to select image subregions of interest for further processing.

2.5 Dealing with imbalanced data

Imbalance in training data can affect network performance by leading to convergence bias towards the majority class. Since imbalanced class samples is common in medical image analysis some strategies such as random over-sampling (ROS) [59], random under-sampling (RUS) [60], dynamic sampling [61], online hard example mining (OHEM) [62], custom loss function [63][64][65], weighted loss [66], custom DNN [67][68], and CNN output thresholding adjustment [69] have been proposed. In order to deal with our rare event detection problem, where LAR frames constitute less then one tenth of the non-LAR

frames, we used both ROS and RUS strategies and created a relatively equal distribution of input training data in each epoch.

2.6 Spatio-temporal networks for classification

Spatio-temporal deep learning networks are designed for learning spatial and temporal features jointly for more accurate prediction. Published deep learning methods involving spatio-temporal information can be categorized into three types: (1) high-dimensional convolutional networks such as 3-D and 4-D convolutional networks; (2) recurrent neural networks (RNNs) such as long short-term memory (LSTM) [70] and gated recurrent unit (GRU) [71] networks; and (3) local spatial features combined with temporal convolutional neural networks (TCN).

Current spatio-temporal solutions that directly use 3-D or even 4-D convolutions [72][73][74] require more hardware memory and computational cost than 1-D and 2-D convolutions. Moreover, for long-term sequential data, memory required for processing temporal information in 3-D and 4-D convolutions increases exponentially as the data becomes longer, which makes it harder to be trained on low-memory GPUs. RNN has been proposed for long-term context-intensive sequential data and applied to many tasks such as arrhythmia detection [75][76], seizure detection [77][78], and action recognition [79][80]. Recently, several spatio-temporal fully convolutional deep learning networks that utilize local spatial features combined with temporal convolutional neural network have been proposed for learning long-term patterns [81][82]. However, due to the low frame rates of flexible transnasal endoscopy and the rare nature of the LAR and occlusion events, state of the vocal folds change momentarily within a few frames. In such rare and short-term cases, use of longer term temporal information hurts rather than helps classification/detection performance. [5] showed that long short-term memory (LSTM) leads to false detections on an empty (all black) video frame when used for vocal fold segmentation despite use of high-speed endoscopy video.

2.7 Segmentation of meningeal microvasculature in confocal microscopy images

While many algorithms have been proposed for segmentation of vascular systems, majority of these works focus on analysis of retinal blood vessels [83][84][85] [86] [87]. These works mostly benefit from publicly available annotated retinal blood vessel data such as DRIVE [88], STARE [89], and ImageRet [90] etc.. Computational analysis of vascular images from confocal microscopy remains limited due to the severe lack of annotated data and challenges associated with 3-D image analysis. Unlike retinal blood vessels, meningeal microvasculature is characterized by irregular shapes, varying scales of vessels, staining and imaging issues.

2.8 Class activate map

For optimum performance, deep learning networks require large amounts of training data. However, data annotation is time-consuming and labor-intensive. Especially in biomedical fields, data is already hard to acquire due to hardware limitations, privacy protection, and saving issues. On top of this, biomedical data annotation becomes even harder due to labor intensive nature of the process and expertise requirements.

Recently, a deep-learning-based technique, named class activation map (CAM), has been proposed to visualize the decision basis in deep learning networks. CAM methods usually generate a rough 2-D class discriminative saliency map for the input image, showing a pixel-wise probability estimation of pixels being used to decide the class label in deep learning classification networks. This visualization improves the interpretability, explainability, and reliability of deep learning networks. Transformation from image-level class labels to pixel-level class activation maps can be used to power weakly-supervised pixel-level tasks such as weakly-supervised object localization and weakly-supervised object segmentation. These tasks can greatly reduce human annotation workload and save time.

Several CAM methods have been proposed, such as gradient-free methods CAM [91] and Ablation-CAM [92], or gradient-based methods such as Grad-CAM [93] and Grad-CAM++ [94]. To generate a saliency map, these methods often combine information from all the channels from a convolutional layer using a weighted sum operation. This process can blend unrelated regions of the target object and affect the energy distribution of the saliency map, lowering the accuracy of downstream tasks such as the weakly-supervised object localization and weakly-supervised object segmentation processes.

2.9 Weakly-supervised object localization and segmentation

Weakly-supervised learning is a branch of machine learning. Weakly-supervised learning recently has received enormous attention, aiming to reduce data annotation workload and maintain the same level of output precision [95] [96]. Based on the types of training data, weakly-supervised learning can be categorized into three types: (1) incomplete supervision, which means only a subset of training data is labeled; (2) inexact supervision, which means that the training data is coarse-grained; (3) inaccurate supervision, which means the training labels are noisy [97]. Based on the purposes, weakly-supervised learning can be divided into different categories according to their desired outputs, such as weakly-supervised object localization (WSOL), weakly-supervised object detection (WSOD), weakly-supervised object classification (WSOC), and weakly-supervised object segmentation (WSOS). Different methods have been proposed to achieve the goal, for example, methods that are based on specific system architectures, feature extraction and refinement, loss function, and training strategies [98].

CHAPTER 3

EVENT DETECTION: DEEPDDK FOR ORAL DIADOCHOKINESIS ANALYSIS

This chapter introduces our 1-D signal event detection deep-learning networks DeepDDK and multi-modal multi-scale DeepDDK (MS-DeepDDK). DeepDDK is a cascade of two 2-D deep-learning networks that use single-scale 1-D signal as input [4]. Multi-modal multi-scale DeepDDK, MS-DeepDDK, is an end-to-end trainable deep-learning network that takes multi-scale 1-D signal as input [28], which is an improvement of our original DeepDDK network [4]. The proposed DeepDDK and MS-DeepDDK are trained with weak supervision. This allows a lightweight deep learning network and enables accurate and fast 1-D signal event detection. The DeepDDK and MS-DeepDDK networks were evaluated on oral diadochokinesis (oral-DDK) audio signals.

In this chapter, we first introduce the background and objective of oral diadochokinesis data analysis. Then, we introduce our proposed DeepDDK and MS-DeepDDK networks, and their corresponding experimental results individually.

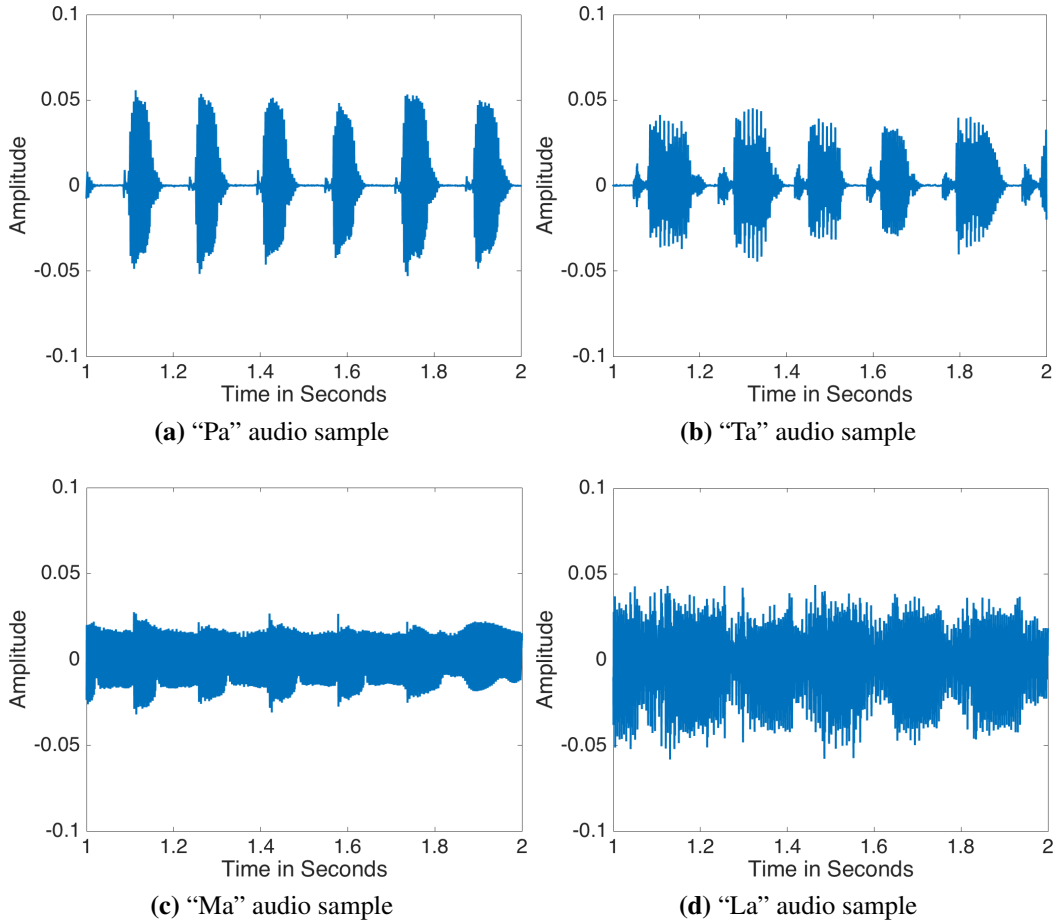
3.1 Introduction

Various neurological disorders such as Parkinson's disease (PD), stroke, amyotrophic lateral sclerosis (ALS), etc. cause oromotor dysfunctions resulting in significant speech and swallowing impairments. Assessment and monitoring of speech disorders offer effective and non-invasive opportunities for differential diagnosis and treatment monitoring of neurological disorders. Oral diadochokinesis (oral-DDK) is a widely used test conducted

by speech-language pathologists (SLPs) to assess speech impairments. Unfortunately, analysis of the oral-DDK tests relies on perceptual judgments by SLPs and are often subjective and qualitative, thus limiting their clinical value. In this chapter, we propose a multimodal oral-DDK test analysis system involving automated processing of complementary 1-D audio and 2-D video signals of both speech and swallowing function. The system aims to automatically generate objective and quantitative measures from the oral-DDK tests to aid early diagnosis and treatment monitoring of neurological disorders. The audio signal analysis component of the proposed system involves a novel multi-scale deep learning network. The video signal analysis component involves tracking mouth and jaw motion during speech tests using our visual landmark tracking software. The proposed system has been evaluated on speech files corresponding to 9 different DDK speech syllables. The experimental results demonstrated promising audio syllable detection performance with an average of 1.6% count error across different types of oral-DDK speech tasks. Moreover, our preliminary results demonstrated added value through combined audio and video signal analysis.

Diagnostic and prognostic accuracy as well as timely intervention and treatment monitoring are important for progressive neurological disorders such as Parkinson's disease (PD), amyotrophic lateral sclerosis (ALS), and multiple sclerosis (MS), since earlier intervention is associated with improved quality of life and survival in these patient populations. Diagnosis and monitoring of neurological disorders involve various medical tests, some of which can be invasive and expensive, prohibiting their effective use. Recent advances in mobile health technologies have lead to the development of non-invasive, more accessible, and affordable new methods and devices not only for diagnosing and monitoring medical conditions, but also for tracking functional decline induced by these diseases. This chapter focuses on development of an oral-diadochokinesis (oral-DDK) analysis software for non-invasive, objective, and quantitative assessment and monitoring of speech disorders common in PD, ALS, MS, and other neurological disorders.

Oral-DDK tasks are universally used by speech-language pathologists (SLPs) for assessment and monitoring of motor speech disorders (e.g., dysarthria and apraxia)[99]. These tasks involve repetitions of single syllables like “Pa”, “Ta”, “Ka”, or sequential multi-syllables such as “Pa-Ta-Ka”, “Buttercup”, etc. as fast as possible, in one breath or within a fixed period of time. SLP use these tasks to estimate diadochokinetic (DDK) rate to provide information about a person’s ability to make rapid speech movements using different parts of the mouth [100]. Manual analysis of DDK rate from audio files is subjective, time intensive, and error-prone. Furthermore, since manual analysis only estimates syllable count, not the locations (timestamps), or production accuracies of the events, rich information that can help diagnosis or monitoring is lost.



(a) “Pa” audio sample

(b) “Ta” audio sample

(c) “Ma” audio sample

(d) “La” audio sample

Figure 3.1. Audio waveform samples for different types of oral-DDK tasks.

In this chapter, we present 1-D signal event detection deep-learning networks DeepDDK and MS-DeepDDK, which take 1-D signal as input and predict start and end timestamps of the detected 1-D signal events. The proposed DeepDDK and MS-DeepDDK can be applied to detect and localize syllables in DDK audio files. The aim of this work is to enable the computation of objective, quantitative measures from the oral-DDK signals to aid early diagnosis and treatment monitoring of neurological disorders.

3.2 DeepDDK Oral Diadochokinesis Analysis Network

Deep learning is a subfield of machine learning that allows learning of high-level abstractions in data through its multi-layer architecture [11]. Inspired by the recent successes of deep learning in speech and image analysis, we have developed DeepDDK, a deep learning based system for automated detection and localization of syllables in oral-DDK tasks.

3.2.1 Network Architecture

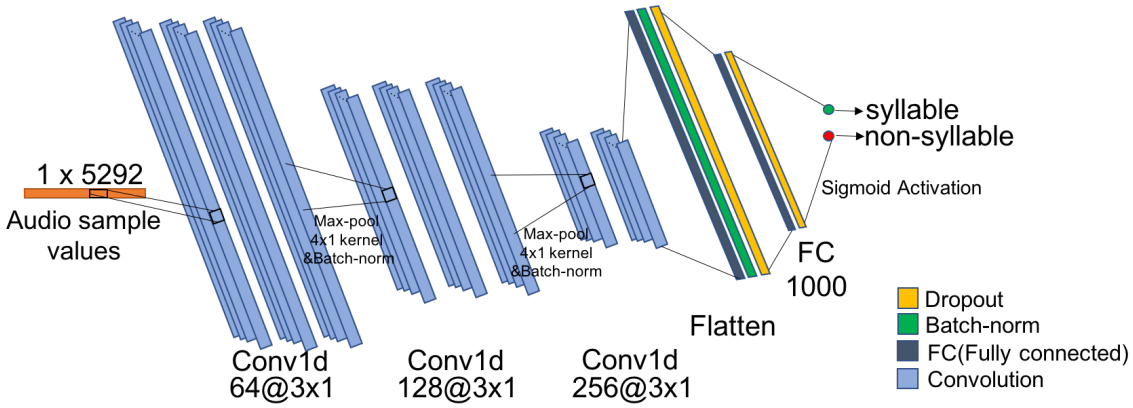
DeepDDK consists of a cascade of two convolutional neural networks (CNNs). The first CNN (CNN-1) segments the 1-D audio signal into syllable vs. non-syllable (silence, breath, etc.) regions. CNN-1 is a classification network that operates on 1-D temporal array of audio samples. Input size is 1×5292, where $5292 = 120ms\times44.1kHz$ corresponds to the product of average syllable duration and sampling rate. Output size is two, corresponding to syllable and non-syllable labels. The CNN-1 network structure is shown in Figure 3.2a and Table 3.1.

The second CNN (CNN-2) locates syllable timestamps within the syllable regions detected by CNN-1. CNN-2 is a 2D regression network that operates on a sequence of audio frames (temporal windows) with length of average event duration $120ms$ (Table 3.2) to predict the precise timestamp of a syllable. For each frame in the input, CNN-2 predicts the probability to contain a syllable. Input size is 15×5292, where 15 is the number of

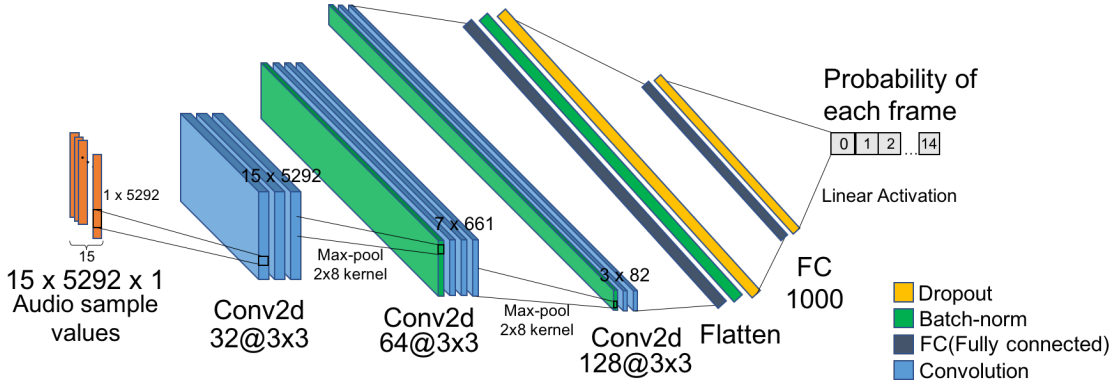
audio frames analyzed and 5292 is the length of a frame as in CNN-1. The CNN-2 network structure is shown in Figure 3.2b and Table 3.1.

3.2.2 Convolutional Neural Networks Training of DeepDDK

Using our custom DDK data collection iOS App, we conducted an IRB-approved study to collect oral-DDK data from seventeen testers for nine tasks (corresponding to syllables "Pa", "Ta", "Ka", "Da", "Ba", "Ga", "La", "Ma", and "Ha"). Following study consent, subjects were instructed to repeat each syllable as fast as they could for 15 seconds. Each task was repeated twice, resulting in 306 audio files of length 15 seconds. All audio files were sampled at 44.1 kHz. Our DeepDDK system relies on availability of labeled training data. In order to label data, we have developed a preliminary unsupervised automated



(a) CNN-1 architecture



(b) CNN-2 architecture

Figure 3.2. CNN-1 and CNN-2 architectures used for DDK syllable detection and localization.

CNN-1: Layer (type)	Size	CNN-2: Layer (type)	Size
input_1 (InputLayer)	(5292, 1)	input_2 (InputLayer)	(15, 5292, 1)
conv1 (Conv1D)	(5292, 64)	conv1 (Conv2D)	(15, 5292, 32)
conv2 (Conv1D)	(5292, 64)	conv2 (Conv2D)	(15, 5292, 32)
conv3 (Conv1D)	(5292, 64)	conv3 (Conv2D)	(15, 5292, 32)
max_pooling1d_1	(1323, 64)	max_pooling2d_1	(7, 661, 32)
batch_norm_1	(1323, 64)	batch_norm_5	(7, 661, 32)
conv4 (Conv1D)	(1323, 128)	conv4 (Conv2D)	(7, 661, 64)
conv5 (Conv1D)	(1323, 128)	conv5 (Conv2D)	(7, 661, 64)
conv6 (Conv1D)	(1323, 128)	conv6 (Conv2D)	(7, 661, 64)
max_pooling1d_2	(330, 128)	max_pooling2d_2	(3, 82, 64)
batch_norm_2	(330, 128)	batch_norm_6	(3, 82, 64)
conv7 (Conv1D)	(330, 256)	conv7 (Conv2D)	(3, 82, 128)
conv8 (Conv1D)	(330, 256)	conv8 (Conv2D)	(3, 82, 128)
flatten_1 (FC)	(84480)	flatten_2 (FC)	(31488)
batch_norm_3	(84480)	batch_norm_7	(31488)
dropout_1	(84480)	dropout_3	(31488)
dense_1 (FC)	(1000)	dense_3 (FC)	(1000)
dropout_2	(1000)	dropout_4	(1000)
dense_2 (FC)	(2)	dense_4 (FC)	(15)
batch_norm_4	(2)	batch_norm_8	(15)
output_classification	(2)	output_regression	(15)

Table 3.1. Layer details for CNN-1 and CNN-2 used for DDK syllable detection and localization.

syllable detector (envelope with local maxima) with a user interface for visualization, navigation, and editing of the results. Results form unsupervised detector were inspected by three experts and corrected according to consensus using our visualization and editing interface. The ground truth consists of a timestamp for each syllable, instead of a region in the audio signal. Locations of these timestamps typically correspond to the sample value maxima in the syllable/event. Considering average event duration (see Table 3.2), 120ms temporal windows (also called 'frames') centered at the ground-truth syllable timestamps are used as positive samples (syllables) to train CNN-1. Each timestamp produces only one positive frame, reducing false positives in the training data compared to use of sliding windows. Shaded regions in Figure 3.3 show these positive samples. As negative samples (non-syllable) to train CNN-1, 120ms temporal windows centered at each midpoint

between two consecutive syllable timestamps are used. In order to prevent information loss, instead of extracting hand-crafted audio features, raw audio sample value is fed into CNN-1. CNN-2, the regression module, aims to predict precise timestamps of the detected syllables. To train the network, 15 sequential frames (one centered at, K centered before, and $15 - K - 1$ centered after the ground truth timestamp, where K is a random number in the range 1 to 13 for robustness) are extracted with a step size of 12ms. Each frame is assigned a score indicating its probability to contain a syllable:

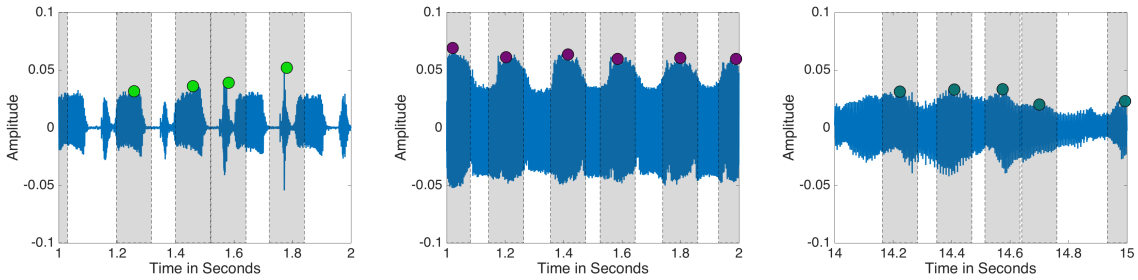
$$P_i = 1 - \frac{|i - i_{GT}|}{15 - 1} \tag{3.1}$$

where i is the index of the specific frame in the sequence and i_{GT} is the index of the frame centered on the ground-truth syllable timestamp.

3.2.3 Convolutional Neural Networks Testing of DeepDDK

DeepDDK syllable detection and localization processes can be summarized as follows. The intermediate outputs from classification network (CNN-1) and regression network (CNN-2) are shown in Figure 3.4b and 3.4c.

Step-1 Classification: Raw audio samples are fed into CNN-1. For each sliding window with stride 12ms, CNN-1 predicts a class label (syllable vs. non-syllable), which is then



(a) "Ta" labeled sample **(b)** "La" labeled sample **(c)** "Ha" labeled sample

Figure 3.3. Sample training data. Colored dots mark ground-truth timestamps, shaded regions mark positive training samples for CNN-1.

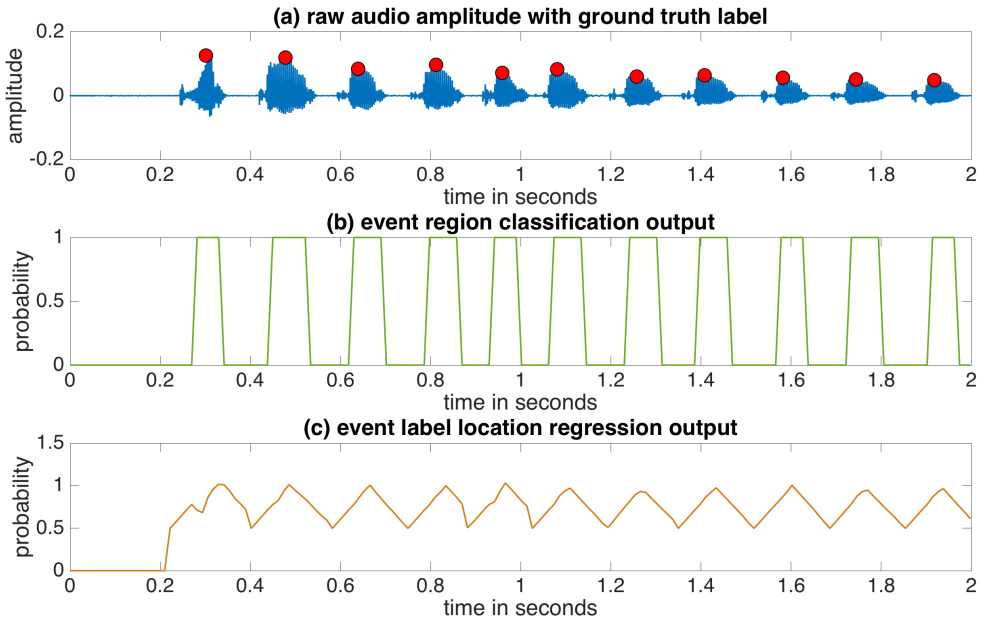


Figure 3.4. Intermediate outputs from the different stages of DeepDDK for a sample "Pa" file. Top panel: original audio signal (blue) with ground-truth timestamps (red). Second panel: output of CNN-1. Third panel: output of CNN-2 where local maxima indicate syllable timestamp.

assigned to the sliding window. The process produces a binary 1D array, $\mathcal{L}$, where $\mathcal{L}(t) = 1$ indicates presence of a syllable at time t.

Step-2 Interval Preprocessing: Morphological closing is applied to $\mathcal{L}$ to fill small gaps in class labels. Syllable event time intervals $\mathcal{E} = \{\mathcal{E}_1, .., \mathcal{E}_n\}$ are identified by applying connected component labeling to $\mathcal{L}$. $\mathcal{E}_i$ represents a region of a syllable. n indicates syllable/event count in the file.

Step-3 Syllable Timestamp Prediction: From each syllable event interval $\mathcal{E}_i$, 15 sequential frames are extracted. If the duration of $\mathcal{E}_i$ is less than 15 frames, the negative frames around $\mathcal{E}_i$ will be included until $\mathcal{E}_i$ duration has 15 frames. Extracted frames are fed into CNN-2 for timestamp score prediction. The center of the frame with the maximum CNN-2 score is marked as the timestamp for event $\mathcal{E}_i$.

3.3 Experimental Results of DeepDDK

As described in Section 3.2.2, we have collected 306 audio files corresponding to 17 subjects, 9 different syllables, and two files for each syllable type. These files were first analyzed by our unsupervised gammatone-based syllable detection software. The detections were then corrected by expert speech pathologists using our visualization and editing interface to produce ground-truth data. Out of these 306 files, 225 files (74%) were used to train the proposed DeepDDK software, and 81 files (26%) were used to test the syllable detection and localization performance. Each audio file was 15sec long. The average number of events per audio file was 74.

We evaluated the system performance in terms of syllable/event count accuracy and syllable/event localization accuracy. Event counting accuracy is evaluated by comparing the number of detected events (DT) to the number of ground truth events (GT). The average event count difference $\frac{1}{N}\sum_{i=1}^{N}|\#DT(i) - \#GT(i)|$ between DeepDDK and ground-truth for $N = 81$ test files is 1.6 events. The average execution time per test file is 1.9s. Figure 3.5 shows detailed, comparative, syllable count accuracy analysis for the proposed DeepDDK and a very recent pre-linguistic speech segmentation tool[1]. DeepDDK results in low syllable count errors and outperforms the pre-linguistic speech segmentation tool[1]. For 81% of the test files, DeepDDK count error is 2 or less ($|\#DT(i)-\#GT(i)| \leq 2$). Considering that the average number of events per file is 74, this corresponds to 2.70% error. For the case of [1], only 72% of the test files have a count error of 2 or less. Figure 3.5 also shows that DeepDDK's highest error for any file is 5, which corresponds to an error of 6.75%, whereas when [1] is used, 17% of the files have a count error higher than 5. We also compared our results with another DDK software from Smekal et al. [2][3], and linear support vector machine (SVM) with Mel-frequency cepstral coefficients (MFCC) features. Table 3.2 presents overall and type-specific event localization performances for DeepDDK. Localization performance is measured in terms of recall ($Recall = \frac{\#TP}{\#GT}$) and precision ($Precision = \frac{\#TP}{\#DT}$) for a given temporal distance threshold. Two temporal distance

thresholds $T = 30ms$ and $T = 60ms$ were used to evaluate performance. If the timestamp of the detected event is located within T milliseconds of the ground-truth event, the detection is considered a true positive detection (TP). We can see in Table 3.2 that event types 'Pa', 'Ka', 'Ba', 'Da', and 'Ga' have very high location accuracies for $T = 30ms$, because of their fairly regular pattern. In contrast, 'Ta' appears to have lower location accuracy. However, this is mostly due to its relatively longer duration (larger than our frame length), which leads to the shift location of the predicted event label.

3.4 MS-DeepDDK Multi-scale Oral Diadochokinesis Analysis Network

The proposed MS-DeepDDK is shown in Figure 6.2. The proposed system consists of two parallel subsystems for improved robustness. The first subsystem is responsible for analysis of audio signals, while the second subsystem is responsible for tracking mouth and

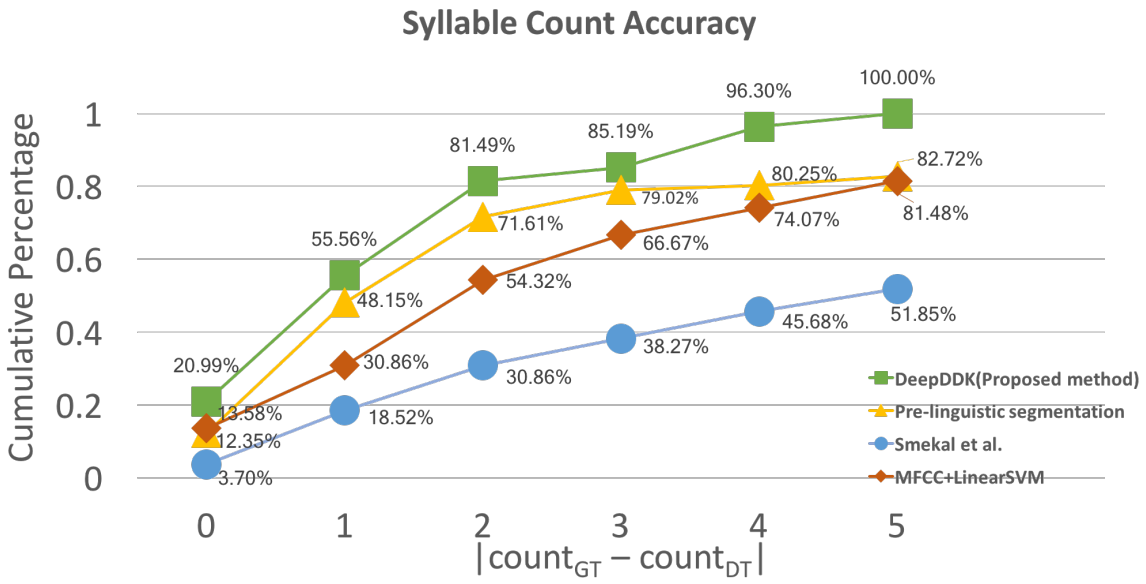


Figure 3.5. Cumulative distribution of event count error for pre-linguistic segmentation[1], Smekal et al.[2][3], MFCC with Linear SVM and our DeepDDK software. Horizontal axis indicates count error (difference between the number of predicted events vs. ground truth events). Vertical axis shows the ratio of the test files. Absolute event count differences of 1, 2, 3, 4, 5 in the graph correspond to percent count errors of 1.35%, 2.70%, 4.05%, 5.40%, 6.75%, respectively (average number of events per file is 74).

Type	Event Duration	Recall		Precision	
		30ms	60ms	30ms	60ms
'Pa'	120ms	0.97	0.98	0.97	0.98
'Ta'	170ms	0.81	0.95	0.81	0.95
'Ka'	140ms	0.91	0.97	0.92	0.98
'Ba'	90ms	0.97	0.98	0.97	0.99
'Da'	130ms	0.89	0.97	0.81	0.98
'Ga'	110ms	0.94	0.96	0.95	0.98
'La'	100ms	0.79	0.90	0.79	0.90
'Ma'	90ms	0.88	0.95	0.89	0.97
'Ha'	140ms	0.85	0.93	0.87	0.95
Average	120ms	0.89	0.95	0.90	0.97

Table 3.2. DeepDDK's location accuracy of different types of syllables.

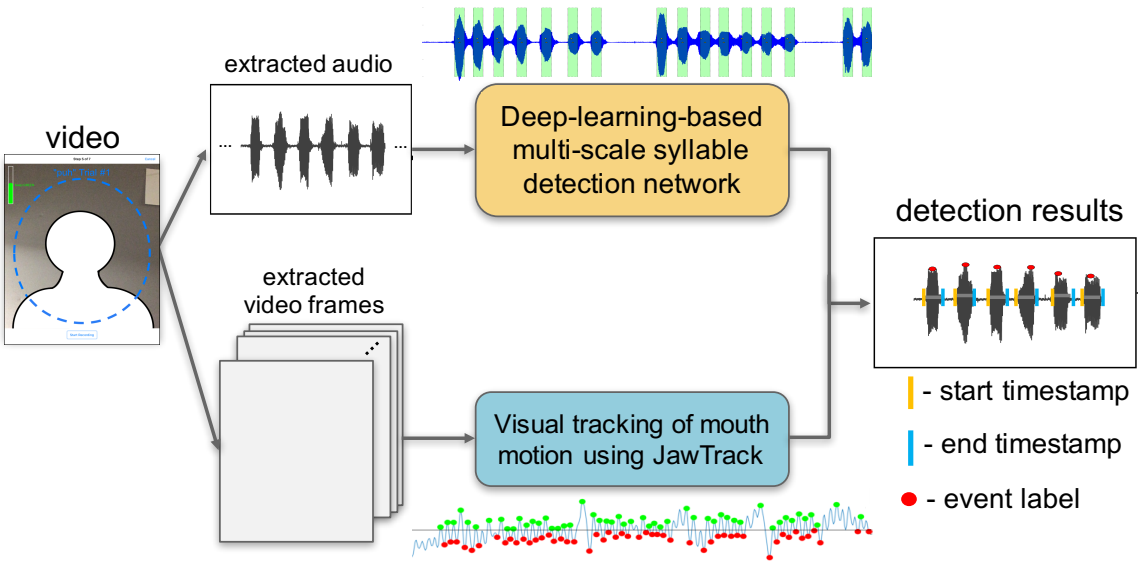


Figure 3.6. The proposed multi-modal and multi-scale oral-DDK analysis pipeline.

jaw motion occurring during oral-DDK tests. Oral-DDK audio signal analysis involves: (1) sliding temporal window syllable/non-syllable classification; and (2) localization of start and end timestamps for individual syllables. During oral-DDK tests, large variations in audio signal amplitude, frequency, and pitch occur due to test subjects' age, gender, fatigue, and severity of their neurological disorders affecting speech production. In order to ensure robust syllable detection despite these signal variations, we propose a multi-scale syllable detection deep learning network.

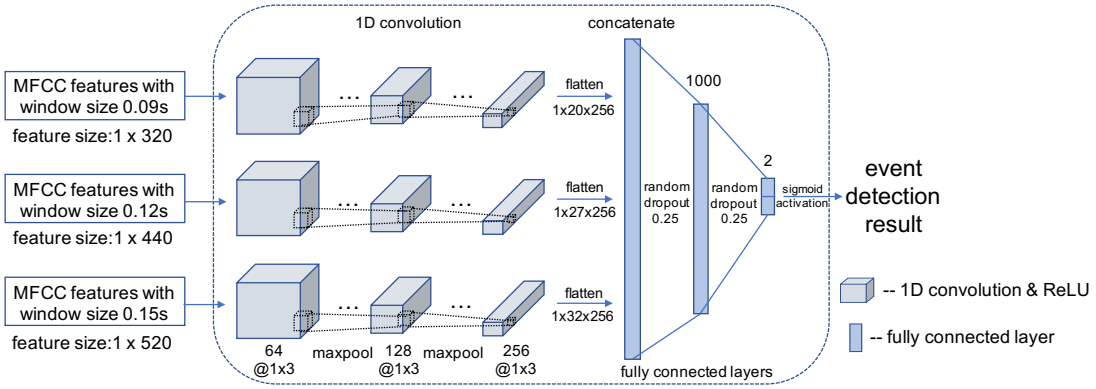


Figure 3.7. The architecture of the proposed multi-scale syllable detection deep learning network MS-DeepDDK.

3.4.0.1 MS-DeepDDK Network Architecture

The proposed multi-scale syllable detection network, MS-DeepDDK, consists of three streams of one dimensional convolutional subnetworks responsible for feature extraction. The inputs to the processing streams are generated using three different temporal scales corresponding to sliding temporal windows of 0.09, 0.12, and 0.15 seconds. These temporal window sizes were selected according to the average syllable durations listed in Table 3.3. In each stream, one-dimensional input vectors are convolved with one-dimensional convolutional kernels. Two max-pooling layers are added between two groups of adjacent convolutional layers to enlarge the receptive field of these convolution operations. The one-dimensional convolutional feature vectors, outputted from the three parallel process-

Table 3.3. Average durations for oral-DDK syllables "Pa" & "Ta" & "Ka" and for various other syllables that are close substitutes.

Average Syllable Durations (seconds)									
"Pa"	"Ta"	"Ka"	"Ba"	"Da"	"Ga"	"La"	"Ma"	"Ha"	**Average**
0.12	0.17	0.14	0.09	0.13	0.11	0.10	0.09	0.14	**0.12**

ing streams, are concatenated and fed to two consecutive fully connected layers to perform syllable/non-syllable classification. The proposed deep learning network architecture is shown in Figure 3.7. Network parameters for the three parallel feature extraction streams are summarized in Table 3.4.

3.4.0.2 Network Training of MS-DeepDDK

Using our custom oral-DDK data collection iOS App, we conducted an IRB-approved study to collect oral-DDK data from 17 testers for 9 tasks (corresponding to syllables "Pa", "Ta", "Ka", "Da", "Ba", "Ga", "La", "Ma", and "Ha"). Following study consent, subjects were instructed to repeat each syllable as fast as they could for 15 seconds. Each task was repeated twice, resulting in 306 audio files of length 15 seconds. All audio files were sampled at 44.1 kHz. Annotated training data was generated by consensus of three domain experts using our interactive oral-DDK annotation software, *TongueTwister*. Ground truth annotations consists of a series of timestamps marking each syllable. Locations of these timestamps typically corresponded to signal local maxima within the syllable/event. Considering average audio event duration (see Table 3.3), three different scales of temporal windows centered at the ground-truth syllable timestamps were used as positive samples (syllables) to train the proposed network. Each timestamp was used to produce three positive temporal windows. Midpoints between consecutive syllable timestamps were used to generate negative samples (non-syllables). As in the case of positive samples, three negative temporal windows were extracted from each midpoint. Mel-frequency cepstral coefficients (MFCCs) [101] were computed for the extracted positive and negative tem-

poral windows. These coefficients were flattened and fed into the proposed network for training.

3.4.0.3 Syllable Detection using MS-DeepDDK

Main processing steps involved in the proposed audio signal based syllable detection and localization processes can be summarized as follows:

1. Classification: MFCC features computed from temporal windows are flattened and fed into the proposed multi-scale network MS-DeepDDK. For each sliding window (with a stride $\frac{1}{10} \times temporal\_window\_size$), the proposed network predicts a class label, syllable versus non-syllable. The process produces a binary 1-D array L, where $L(t) = 1$ indicates the presence of a syllable at time t.

2. Interval preprocessing: Morphological closing is applied to L to fill small gaps in class labels. Syllable event time intervals $E = \{E1, .., En\}$ are identified by applying connected component labeling to L. E_i represents the interval of time corresponding to a syllable. n indicates the syllable/event count in the file.

3. Syllable timestamp prediction: For each syllable event interval E_i, the center of the interval is extracted as the final timestamp prediction for each syllable event.

3.4.1 Visual Oral-DDK Analysis through Mouth and Jaw Motion Tracking

For visual analysis of mouth and jaw motion during oral-DDK tests, we use JawTrack tracking software [102][103]. JawTrack is our group's visual tracking software that provides functionalities of landmark tracking, automated and manual motion event detection, and automated computation of biologically motivated outcome measures. JawTrack uses normalized 2-D cross correlation [104] and motion pattern analysis to track landmarks of interests selected by users. JawTrack supports multiple image modalities (i.e., visible and X-ray) and has been used for tracking landmarks of interest on various animal (mice, rat, dog) and human motion studies [103]. In order to track mouth and jaw motion during

oral-DDK tests, we have marked three locations of interest (i.e., philtrum, mentolabial sulcus, and chin) with colored stickers applied to the skin. The three markers are tracked, visualized, and analyzed using JawTrack. The philtrum is used as a reference point. Distances of mentolabial sulcus and chin to philtrum in time are used to quantify motion and open/closed states of the mouth. A sample screenshot of the JawTrack interface during visual DDK analysis is shown in Figure 3.8. In the example, 1-D plot illustrates the distance between the philtrum and mentolabial sulcus in time, where the green and red dots on the graph mark the automatically detected opened and closed states of the mouth/jaw.

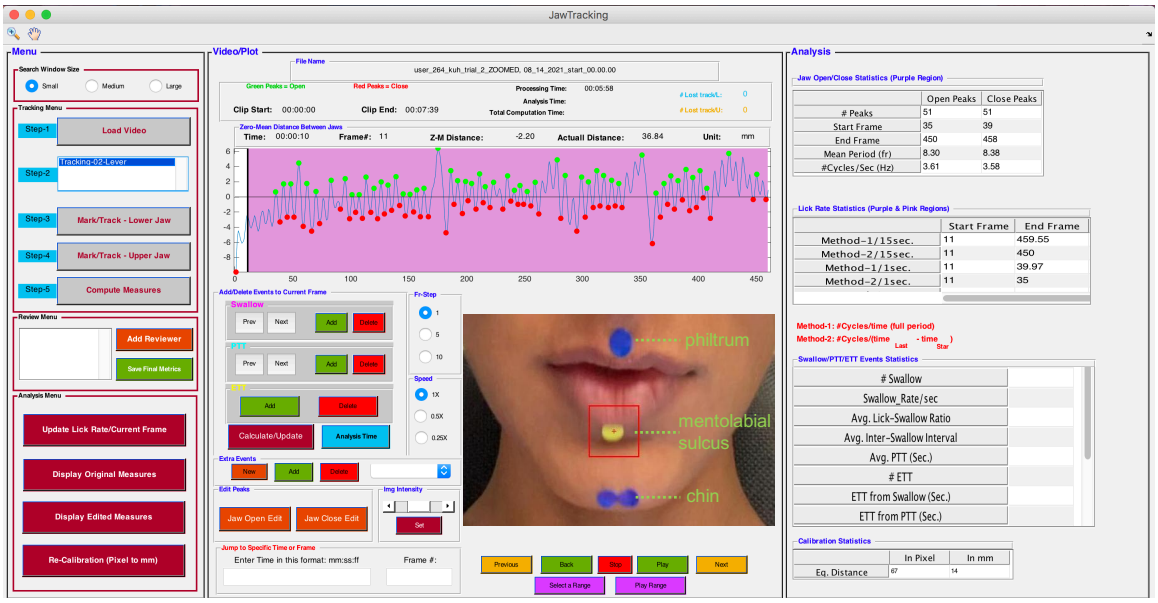


Figure 3.8. Sample screenshot for JawTrack visual tracking software during oral-DDK mouth and jaw motion analysis.

3.5 Experimental Results of MS-DeepDDK

3.5.1 Oral Syllable Detection Accuracy

As described in Section 3.4.0.2, we have collected 306 audio files from 17 subjects, corresponding to 9 different syllables, and two audio files for each syllable type. These files were first analyzed by unsupervised gammatone-based [105] syllable detection option

Table 3.4. Network parameters for the feature extraction streams in the proposed MS-DeepDDK network.

Layer name	kernel size, # channels
conv1	1 x 3, 64
conv2	1 x 3, 64
conv3	1 x 3, 64
maxpooling1D	1 x 4
conv4	1 x 3, 128
conv5	1 x 3, 128
conv6	1 x 3, 128
maxpooling1D	1 x 4
conv7	1 x 3, 256
conv8	1 x 3, 256

in our TongueTwister software. The detections were then corrected by expert SLPs using our visualization and editing interface to produce ground-truth data. Out of these 306 audio files, 225 audio files (74%) were randomly selected for training the proposed multi-scale DDK syllable detection network MS-DeepDDK. The remaining 81 files (26%) were used to test the syllable detection performance. Each audio file was 15 seconds long. The average number of events per audio file was 74.

We compare the performance of the proposed MS-DeepDDK network to four DDK syllable detection approaches: (1) pre-linguistic audio segmentation [1]; (2) non-linear dynamic features combined with empirical mode decomposition (EMD); [3, 2], (3) Mel-frequency cepstral coefficients (MFCC) combined with linear support vector machines (SVM); and (4) our earlier deep DDK syllable detection network DeepDDK [4].

Evaluations were carried out in terms of Mean Absolute Error (MAE), Mean Square Error (MSE), and Cumulative Score (CS) [106, 107] that compared the number of syllable events in the ground truth versus in the automated predictions. These evaluation measures are defined as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|g_i - p_i| \tag{3.2}$$

Table 3.5. Mean absolute error (MAE) and mean square error (MSE) comparisons for oral-DDK syllable detection.

	MAE	**MSE**
Pre-linguistic [1]	3.2375	31.6375
MFCC+linearSVM	3.5432	31.0741
Smekal et al. [2] [3]	17.6420	4178.3580
DeepDDK [4]	1.6049	4.4198
MS-DeepDDK (proposed)	**1.1852**	**2.9630**

Table 3.6. Cumulative scores (CS) for the compared methods for different count error tolerances.

	CS(0)	**CS(1)**	**CS(2)**	**CS(3)**	**CS(4)**	**CS(5)**
Smekal et al. [2] [3]	0.0370	0.1852	0.3086	0.3827	0.4568	0.5185
MFCC+LinearSVM	0.1358	0.3086	0.5432	0.6667	0.7407	0.8148
Pre-linguistic segmentation [1]	0.1235	0.4815	0.7161	0.7902	0.8025	0.8272
DeepDDK [4]	0.2099	0.5556	0.8149	0.8519	0.9630	**1.0**
MS-DeepDDK (proposed)	**0.3210**	**0.7407**	**0.8642**	**0.9136**	**0.9753**	**1.0**

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(g_i - p_i)^2 \tag{3.3}$$

$$CS(x) = \frac{N_{e\leq x}}{N} \times 100\% \tag{3.4}$$

where g_i and p_i represent the ground truth and predicted number of syllable events in oral-DDK audio file i, N is the total number of test audio files, and $N_{e\leq x}$ is the number of files where the prediction makes an absolute count error of less than or equal to error tolerance x.

Syllable detection mean absolute errors (MAE) and mean square errors (MSE) are listed in Table 3.5. The proposed multi-scale oral-DDK syllable detection network MS-DeepDDK achieves the lowest MAE and MSE errors among all five algorithms with an average absolute count error of less than 1.2, corresponding to less than 1.60% error rate, given that the average total number of syllable events in a file is 74. MS-DeepDDK im-

Table 3.7. Detailed cumulative score (CS) analysis for the proposed MS-DeepDDK network using five-folds cross-validation.

Fold	Cumulative score (CS) of different counts error tolerance						
	0	1	2	3	4	5	>5
1	44.44%	81.48%	90.74%	96.30%	98.15%	100%	0.00%
2	33.33%	68.52%	85.19%	94.44%	96.30%	96.30%	3.70%
3	33.33%	76.39%	88.89%	91.67%	95.83%	97.22%	2.78%
4	22.22%	57.41%	75.93%	87.04%	94.44%	96.30%	3.70%
5	38.89%	73.61%	84.72%	97.22%	100%	100%	0.00%
Avg.	34.44%	71.48%	85.09%	93.33%	96.94%	97.96%	2.04%

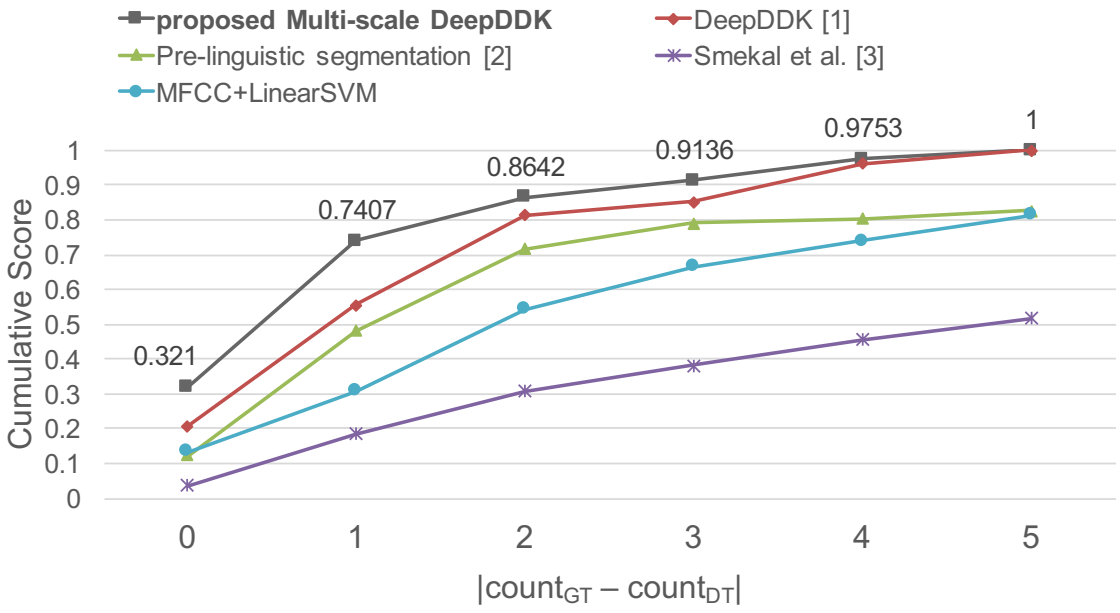


Figure 3.9. Cumulative scores (CS) for different count error tolerances. Comparison between methods: pre-linguistic segmentation [1], Smekal et al. [2, 3], MFCC with Linear SVM, DeepDDK [4], and the proposed multi-scale DeepDDK. Absolute event count differences of 0, 1, 2, 3, 4, 5 in the x-axis of the graph correspond to percent count errors of 0%, 1.35%, 2.70%, 4.05%, 5.41%, 6.76%, respectively (using the average number of events per file as 74).

proves MAE and MSE scores with respect to the next best result DeepDDK [4] by 26.15% and 32.96% respectively.

For a more detailed analysis of the syllable detection errors, we compute and compare cumulative score (CS) measures for absolute count error tolerances of 0, 1, 2, 3, 4, and 5

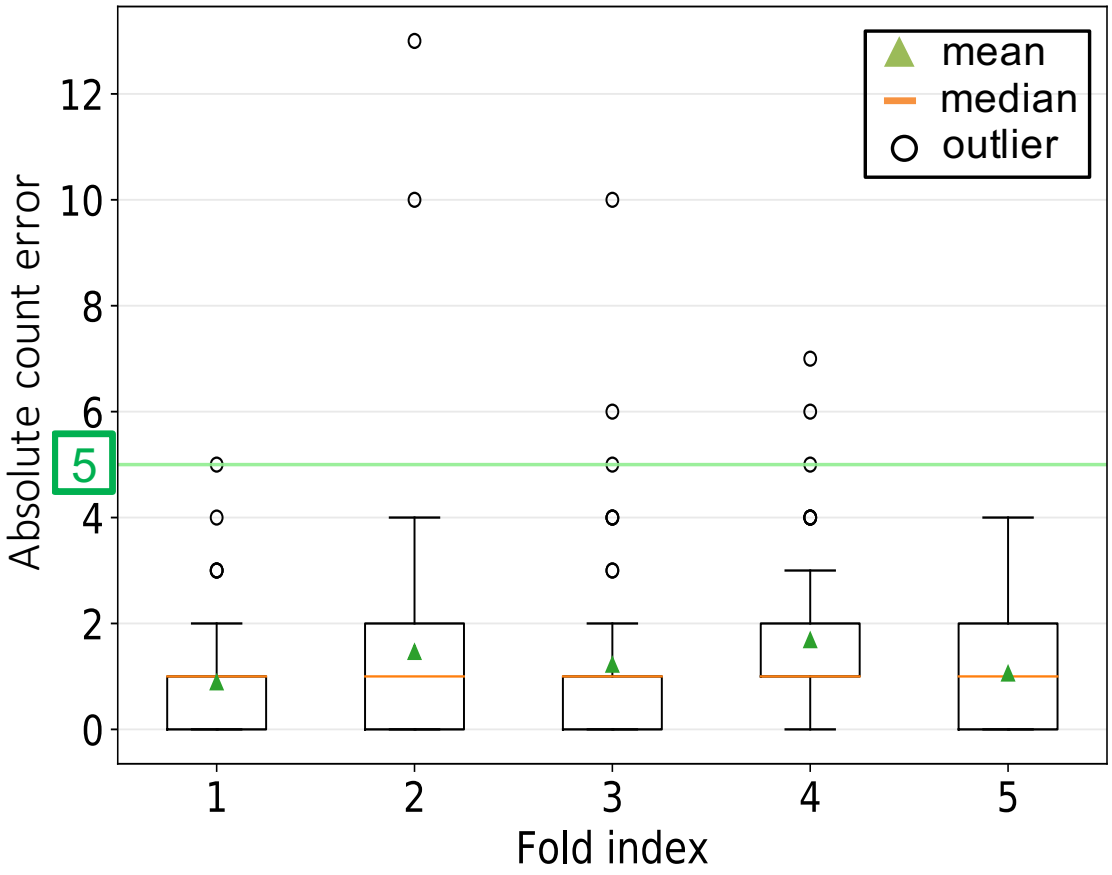


Figure 3.10. Boxplot of the absolute count errors for five-folds cross-validation.

corresponding to percent error tolerances of 0%, 1.35%, 2.70%, 4.05%, 5.41%, and 6.76% respectively. Cumulative score plots for the compared approaches are shown in Figure 3.9. Detailed cumulative scores for different tolerance levels are listed in Table 3.6. MS-DeepDDK outperforms the compared approaches for all error tolerances and outperforms the next best approach DeepDDK [4] by 11.11%, 18.51%, 4.93%, 6.17%, and 1.23% for absolute count error tolerances of 0 to 4, respectively.

3.5.2 K-folds Cross-validation of Syllable Event Detection of MS-DeepDDK

We further analyzed the performance of the proposed MS-DeepDDK approach using five-folds cross-validation. All oral-DDK audio files in the study, 306 files in total, are

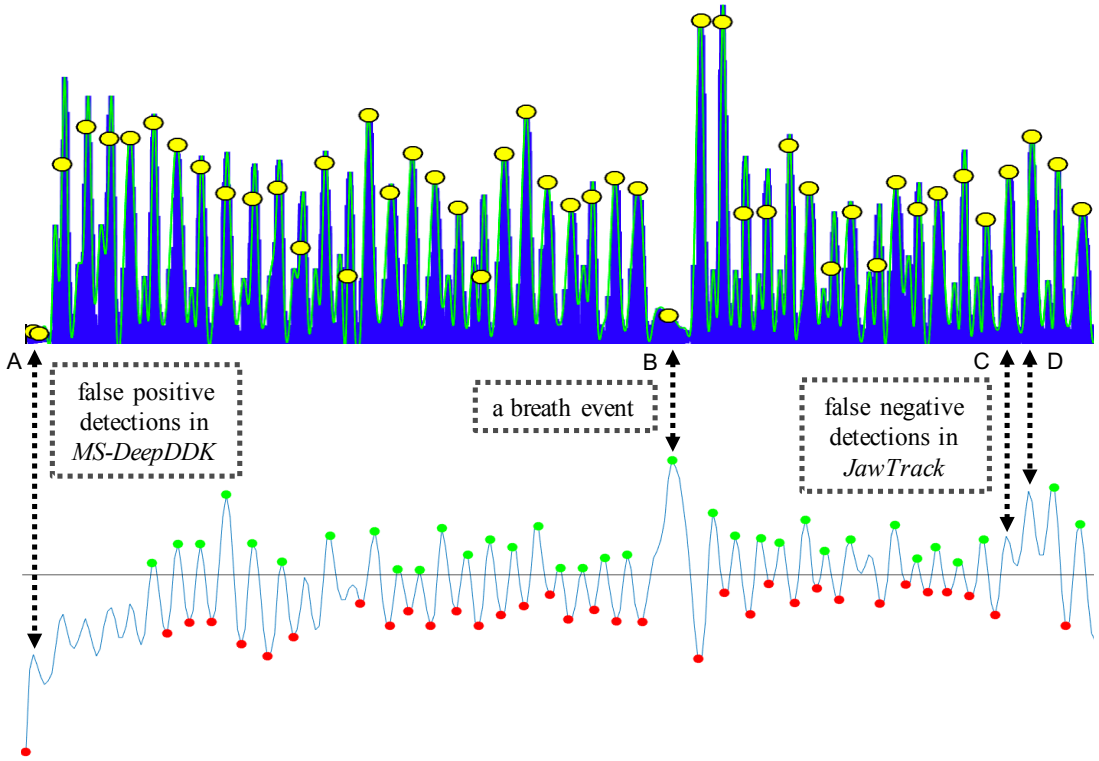


Figure 3.11. Audio-visual analysis of oral-DDK tests. Top: The blue waveform is the original audio waveform. The green line on top of the blue waveform is the signal envelope. Yellow dots mark event timestamps generated by the proposed MS-DeepDDK network. Bottom: the blue signal denotes the distance between the philtrum and mentolabial sulcus in time, automatically computed using our JawTrack visual tracking software. Green and red dots represent opened and closed states of the mouth/jaw outputted by JawTrack.

divided into five folds. Five experiments are performed where one fold of files is kept for testing and the remaining four folds of files are used for training of the proposed network MS-DeepDDK. The five-folds cross-validation results are shown in Table 3.7. As we can see on the last row in Table 3.7, over 85% of the test audio files have an absolute count error equal to or less than 2, corresponding to a 2.7% error rate given that the average number of events per audio was 74. On average, nearly 98% of the test audio files have an absolute count error equal to or less than 5, corresponding to a 6.8% error rate given that the average number of events per audio was 74.

Boxplots of the five-folds cross-validation results are shown in Figure 3.10. Mean and median values of each fold are less than 2 absolute count errors (2.7% error rate), demonstrating robustness across the whole dataset. There are six audio files that have an absolute count error larger than 5, three of which are "La" audio files, one is "Ga", and one is "Ha". These syllables have less clear phonetic boundaries between two adjacent syllables and irregular audio waveforms compared to standard oral-DDK test syllables "Pa", "Ta", "Ka".

3.5.3 Visual Oral-DDK Analysis

Three common sources of errors during oral-DDK audio signal analysis are: (1) background noise in the environment, (2) non-syllable events such as breaths or coughs that produce sound, and (3) weak audio signal amplitudes particularly due to fatigue. Visual cues complement audio signals to overcome these error sources.

Figure 3.11 illustrates the complementary nature of the audio-based MS-DeepDDK and visual JawTrack oral-DDK event detection outputs. Vertical lines marked with labels A, B, C, and D illustrate specific cases where use of both signals improve detection accuracy. (A) Weak false detections by MS-DeepDDK can be filtered-out using JawTrack output that does not detect mouth/jaw opening motion at those timestamps. These false-positive detections can be caused by background noise. (B) Weak false detection by MS-DeepDDK corresponding to a breath can be filtered-out using JawTrack output, where breath events are clearly characterized by abnormally larger signal peaks in height and width (motion amplitude and duration). (C-D) Missed detections in mouth/jaw motion analysis can be corrected using clear syllable signatures in MS-DeepDDK.

3.6 Conclusion

We have presented DeepDDK and MS-DeepDDK, which are deep-learning-based systems for automated analysis of oral-DDK tasks. The proposed system allows objective

and quantitative analysis of oral-DDK data corresponding to a task routinely used by SLPs for assessment and monitoring of oral motor speech abilities. Experimental results show robust syllable detection and localization capabilities across different types of DDK tests. Use of complementary audio-visual cues leads to further robustness. Accurate, objective, quantitative analysis of oral-DDK data is of great significance because these tests can be potentially used to facilitate diagnosis and monitoring of neurological disorders, particularly progressive ones such as PD, ALS, and multiple system atrophy. Our future plans include improved fusion of audio and visual cues and testing of the proposed system for analysis of early Parkinson's disease patient data.

Acknowledgement

This work was partially supported by an award from the University of Missouri School of Medicine TRIUMPH (Translational Research Impacting Useful and Meaningful Precision Health) Initiative. We graciously thank Maria Martinez (medical student) for assistance with video recordings of test subjects.

CHAPTER 4

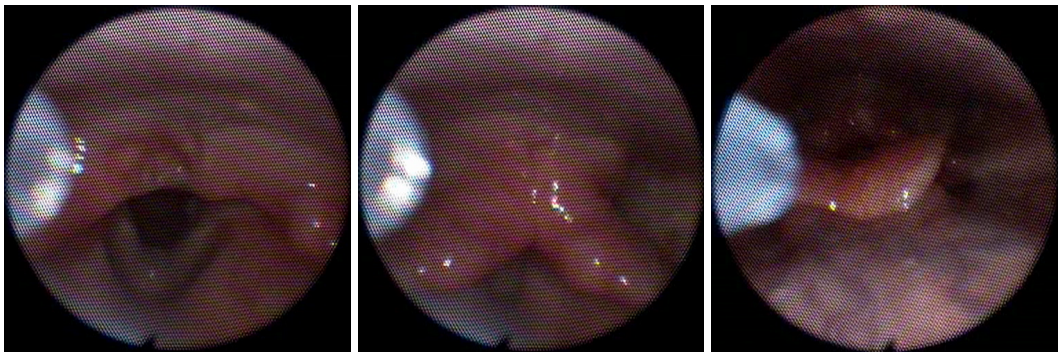
VIDEO SEGMENTATION: ORTHOGONAL REGION SELECTION AND LARNET-STC NETWORKS

This chapter introduces spatial-temporal orthogonal region selection (ORS) networks LARNet and LARNet-STC for 2-D + time (video) data analysis. The LARNet [25] is a two-stream deep classification network using 2-D spatial information, which combines the novel orthogonal region selection (ORS) attention that improves the overall classification result. The LARNet-STC [30] is a spatial-temporal orthogonal region selection deep classification network, which takes sequential 2-D images as input. The LARNet-STC is a further improvement of LARNet using temporal information. The proposed two deep learning networks are end-to-end trainable. The proposed LARNet and LARNet-STC can be applied to 2-D image-based video event detection. We applied our proposed LARNet and LARNet-STC to the human trans-nasal laryngeal endoscopy video for extracting objective and quantitative results for laryngeal adductor reflex (LAR) analysis.

In this chapter, we will first introduce the background and objective of the human trans-nasal laryngeal endoscopy video analysis. Then, we will introduce our proposed LARNet and LARNet-STC and their corresponding experimental results respectively. Finally, we will conclude this chapter.

4.1 Introduction

Vocal folds (VFs), which are also called vocal cords, are two bands of soft muscles located at the larynx (voice box), the tubular structure that connects the throat to the trachea



(a) open VFs **(b)** closed VFs **(c)** visual occlusion

Figure 4.1. Sample images for the three vocal fold state classes.

(windpipe). VFs operate like a valve in the upper airway, opening and closing mainly innervated by the recurrent laryngeal nerve to coordinate breathing, swallowing, and speaking [108] [109]. When speaking, the VFs vibrate and allow air to pass from the lungs through the cords to produce human voice. As a crucial component of the airway, VF dysfunction can lead to breathing difficulty (dyspnea) and swallowing dysfunction (dysphagia) that can significantly endanger a patient's life, or voice impairment (dysphonia) that can affect the quality of life [110]. VF dysfunction occurs when the VFs can't open or close correctly. Inappropriate VF closure can impede breathing or speaking, whereas improper VF opening can allow food and liquid to be inhaled into the airway, causing choking and/or lung infection (aspiration pneumonia) [111]. While numerous medical conditions have been shown to result in life-threatening VF dysfunction, the most prevalent triggers are neurological disorders (e.g., stroke, Parkinson's disease, and amyotrophic lateral sclerosis) and head and neck cancer [112]. The major morbidity and mortality in these conditions or diseases are caused by aspiration pneumonia [113], thus emphasizing the clinical needs for advanced medical management of VF dysfunction.

Flexible Endoscopic Evaluation of Swallowing with Sensory Testing (FEESST) is a clinical test conducted by speech-language pathologists (SLPs) and otolaryngologists to

closely inspect the motor and sensory functions of the VFs and to assess the risk of aspiration [114]. During the FEESST procedure, a thin flexible endoscope is inserted through the nose into the larynx to visualize the VFs, which can be known as nasal laryngoscopy. During the procedure, small puffs of air are delivered to the VFs through the working channel of the endoscope. These puffs of air stimulate the laryngeal mucosa near the VFs and trigger a laryngeal adductor reflex (LAR) where the VFs abruptly close momentarily (less than 1 second) as an airway protection reflex to prevent accidental invasion of "foreign" materials into the lungs. This airway protection is innervated by the vagus nerve, which is located at the throat between the top of the vocal folds to the tip of the epiglottis. While FEESST is a routinely conducted test in clinical practice, the generated laryngoscopy videos (if they are even recorded) are only visually inspected, resulting in loss of valuable clinical information that can potentially be used to help early diagnosis, plan treatment options, and monitor disease progress and treatment effectiveness. In contrast to visual inspection that only checks the occurrence of a LAR event, video analysis can objectively quantify the duration of LAR events and detect subtle VF dysfunction, which may easily be neglected by visual inspection. Our group has developed a patented air pulse device and method to reliably evoke the LAR and visualize the entire larynx during endoscopic LAR testing [115]. This chapter presents a deep learning-based automated video analysis system for automated detection of laryngeal adductor reflex (LAR) events in endoscopic LAR testing to enable objective, quantitative analysis of VF function. The ultimate aims for the proposed system are objective and quantitative monitoring of disease progression and treatment response; and generation of novel quantitative data to facilitate development of data-driven preventative strategies for life-threatening diseases like aspiration pneumonia caused by VF dysfunction.

Recent studies demonstrated the scientific and clinical utility of quantitative vocal fold motion analysis in prediction and monitoring of dysphagia-related aspiration pneumonia disease [116]; in monitoring dysfunction after recurrent laryngeal nerve (RLN) injury and

during the RLN regeneration [46][117]. Advances in artificial intelligence, particularly in deep learning, have started to bring promising approaches and results to biomedical image analysis, facilitating automated and quantitative analysis for various applications such as microorganism counting [118], breast cancer image analysis [119], automated segmentation for lung cancer radiotherapy [120], MRI and CT bladder segmentation [121], and COVID-19 medical image analysis [122].

In this chapter, we propose a spatio-temporal deep learning network to identify VF states open, closed, and occluded, in order to detect and quantify LAR events in laryngeal endoscopy videos. This is a challenging task because of numerous factors such as anatomical variations in different patients; diverse illumination conditions such as over/under-exposure, artifacts such as glare, camera focus problems, motion blur caused by the endoscope operation; scale variations due to anatomy and position of the endoscope with respect to the VFs; camera or patient motion; partial or full visual obstruction of the VFs due to camera fogging, saliva accumulation, camera position, and motion of the surrounding anatomical structures; the limited number of training videos, and highly imbalanced data due to the rare nature of the LAR events. The distribution of the various challenging cases within the whole dataset and test dataset are listed in Table 4.1.

Table 4.1. Distribution of the various challenging cases in the dataset.

Problem	number of images (percentage)	
	whole dataset	testing set
Out-of-focus	440 (4.83%)	342 (15.04%)
Over-exposure	963 (10.58%)	406 (17.85%)
Low-light	752 (8.26%)	261 (11.48%)
Out of region of interest	531 (5.83%)	134 (5.89%)
Off-center	2588 (28.43%)	418 (18.38%)

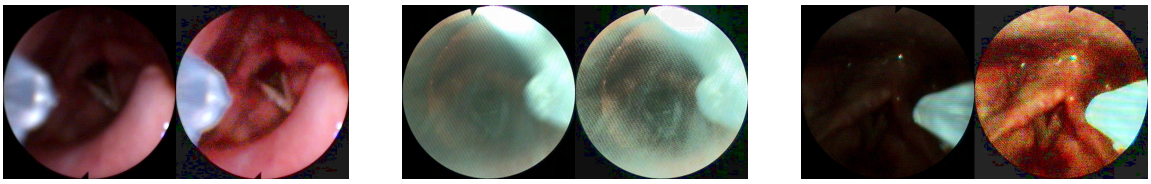
Figure 4.2 shows some sample laryngoscopy frames illustrating these challenging cases.

The proposed network classifies each video frame into one of three classes: non-LAR (open VFs), LAR (closed VFs), and visual occlusion (where the VFs are either obstructed by other anatomical structures or are out of the endoscope camera field of view). Sample video frames of these three different classes are shown in Figure 4.1. The proposed network is designed to directly classify a video frame into VFs' open, closed, or occluded states, without first segmenting or tracking the VFs to estimate their states. This direct, one step process drastically reduces the manual annotation workload to generate network training data from labor-intensive VF segmentation or motion tracking to rapid frame-level class label assignment. Furthermore, as we demonstrate later, direct LAR detection through image classification outperforms segmentation-based LAR detection performance.

The proposed approach incorporates a novel orthogonal region selection (ORS) subnetwork that combines global and local image information. ORS functions like an unsupervised spatial attention mechanism, while the video context incorporates temporal information to improve VFs state estimation accuracy. The temporal context information is extracted from five sequential video frames centered at the target video frame. We call the sliding window of sequential video frames an "image block." Besides the VFs state classification task, our proposed ORS subnetwork can be used along with other deep classification networks as a spatial attention mechanism to strengthen feature extraction. To the best of our knowledge, we are the first group that proposed automated video analytics solutions for the endoscopic analysis of the LAR tests and this work is the first direct VF state estimation system utilizing spatio-temporal context. The proposed network augments the image features extracted from the target frame with temporal context from the neighboring video frames, to improve the classification accuracy of the target frame.

The contributions of this chapter are summarized below:

1. We propose a novel orthogonal region selection (ORS) network that combines local and global image information for the task of VF state classification. This proposed



(a) out-of-focus sample **(b)** over-exposured sample **(c)** low-light sample

Figure 4.2. Sample laryngoscopy video frames illustrating different processing challenges. (a-c) Left images show original video frames, right images show corresponding histogram equalized images.

method directly maps its input to a VF state without prior generation of segmentation masks, regions of interest, or motion trajectories, which reduces manual annotation.

2. We introduce an end-to-end trainable spatio-temporal network that integrates temporal context with orthogonal region selection to further improve the classification accuracy. To the best of our knowledge, this is the first spatio-temporal deep classification network for analysis of flexible transnasal endoscopy videos.

3. The proposed orthogonal region selection (ORS) subnetwork can be applied on top of other deep learning classification networks as a spatial attention mechanism to strengthen feature extraction.

4. This proposed deep learning system generates promising performance in detecting rare LAR and occlusion events with limited amounts of training data in general and with highly imbalanced data between non-LAR class and other classes.

4.2 Related Work

4.2.1 Laryngeal Endoscopy Video Analysis

Majority of the earlier works on laryngeal endoscopy video analysis (i.e., for VF segmentation [41], for analysis of VF vibration patterns [42][43], or for analysis of VF shape and vascular defects) were limited to processing of high-speed rigid transoral (through the

mouth) laryngoscopy videos. Compared to these videos that have a higher frame rate, higher image resolution, and better image quality, this chapter focuses on processing of flexible transnasal (through the nose) endoscopy videos widely used in clinical practice that lack these advantages. The earlier work have also mostly relied on handcrafted/engineered features which are not very robust and require parameter tuning. Recently, various deep learning approaches have been proposed for analysis of laryngeal videos. In [6], we have developed a deep convolutional regression network for segmentation of the glottal region. Annotated training data for this network was generated by extending our previous interactive VF tracking software [46]. In [47], a cascade of two networks was proposed to segment the laryngeal structures. In [5], an U-Net [26] based network has been developed for glottis region segmentation in high-speed endoscopic videos. The results from the network is further improved by utilizing Long Short-Term Memory (LSTM) [48] blocks to incorporate the temporal context.

While great improvement over the previous handcrafted/engineered works, those networks that are designed for VF segmentation are not very suitable for LAR event detection for four main reasons: (1) segmentation networks require ground truth segmentation masks for training. Preparation of these training masks is a labor intensive and time-consuming task requiring domain expertise. Whereas frame level single class label assignment required for our proposed system is much more efficient. (2) Segmentation networks are often tested on video frames where VFs are fully visible. They often detect spurious regions when VFs are occluded or are not in the field of view of the endoscope. (3) Similarly, LAR events lead to tracking failures due to abrupt motion. (4) Additionally, most of the earlier work on VF analysis (except [5]) did not incorporate temporal video context and analyzed individual video frames independently leading to loss of valuable temporal information. In this chapter, we perform explicit VF state estimation and incorporate temporal video contextual information for robust state estimation. Beside LAR event detection, the proposed network can be used as a preprocessing step for VF segmentation or VF tracking

to determine temporal intervals of interests for further processing. The following subsections review related work specific to the challenges we are addressing in this chapter.

4.2.2 Objects in Different Scales

In the deep convolutional neural networks, max-pooling layers that are designed to enlarge the convolution field of view play a critical role. They present a compromise between the hardware capabilities and the convolutional kernel size. They help a network extract and learn abstract and transformation invariant features from an image. Unfortunately, this process can also lead to elimination of small objects. If these small objects are of interest to the given task, network performance can get adversely affected. In this application, VFs typically occupy a small portion (typically $< 25\%$) of the endoscope field of view. In some cases, when the endoscope is positioned further away from the VFs, or when the endoscope is being pulled out of the larynx, the apparent size of the VFs can be even smaller. Furthermore, the relatively far distance between the VFs and the endoscope makes the camera harder to autofocus. Depending on the position and orientation of the endoscope, VFs can appear at different positions in a video frame. VF state estimation performance is expected to be adversely affected by the small size of VFs in these inspected images.

To deal with the small object classification/segmentation problems, regions of interest are usually extracted from the input image before performing classification or segmentation tasks. For example, deep object detection networks such as Fast-RCNN [123], Mask-RCNN [124], feature pyramid [50], etc. first propose a set of bounding boxes before performing further analysis. These detection networks require ground truth bounding boxes (or segmentation masks) besides class labels for training, which drastically increase the manual annotation workload. In this chapter, we propose an unsupervised region selection scheme (orthogonal region selection (ORS) subnetwork) that selects an image subregion without need for manual annotation.

4.2.3 Orthogonality

Orthogonality has been used for network initialization in deep neural networks (DNNs) [51][52]. Orthogonal weight normalization has been proposed to improve network generalization capabilities [53]. There is a growing interest on use of deep neural network (DNN) layer weight pruning techniques to reduce feature redundancy in DNNs. The aim is to improve the generalization and precision capabilities of DNNs [54]. However, the popularization of the weight pruning technique is limited by its difficulty of implementation. Orthogonality has also been used in network initialization and regularization to prevent gradient vanishing or exploding problem in training very deep neural networks [55][56]. [57] proposes a loss function that encourages the features of different classes to be orthogonal to each other. Orthogonal deep features decomposition has been proposed to improve face recognition accuracy [58]. In this chapter, we rely on feature orthogonality to select the subregion of interest for further processing.

4.2.4 Imbalanced Data

Imbalance in training data can affect network performance by leading to convergence bias towards the majority class. Since imbalanced class samples is common in medical image analysis some strategies such as random over-sampling (ROS) [59], random under-sampling (RUS) [60], dynamic sampling [61], online hard example mining (OHEM) [62], custom loss function [63][64][65], weighted loss [66], custom DNN [67][68], and CNN output thresholding adjustment [69] have been proposed. In order to deal with our rare event detection problem, where LAR frames constitute less then one tenth of the non-LAR frames, we used both ROS and RUS strategies and created a relatively equal distribution of input training data in each epoch.

4.2.5 Spatio-temporal Networks for Classification

Spatio-temporal deep learning networks are designed for learning spatial and temporal features jointly for more accurate prediction. Published deep learning methods involv-

ing spatio-temporal information can be categorized into three types: (1) high-dimensional convolutional networks such as 3D and 4D convolutional networks; (2) recurrent neural networks (RNNs) such as long short-term memory (LSTM) [70] and gated recurrent unit (GRU) [71] networks; and (3) local spatial features combined with temporal convolutional neural networks (TCN).

Current spatio-temporal solutions that directly use 3D or even 4D convolutions [72][73][74] require more hardware memory and computational cost than 1-D and 2-D convolutions. Moreover, for long-term sequential data, the memory required for processing temporal information in 3D and 4D convolutions increases exponentially as the data becomes longer, which makes it harder to be trained on low-memory GPUs. RNN has been proposed for long-term context-intensive sequential data and applied to many tasks such as arrhythmia detection [75][76], seizure detection [77][78], and action recognition [79][80]. Recently, several spatio-temporal fully convolutional deep learning networks that utilize local spatial features combined with temporal convolutional neural network have been proposed for learning long-term patterns [81][82]. However, due to the low frame rates of flexible transnasal endoscopy and the rare nature of the LAR and occlusion events, state of the vocal folds change momentarily within a few frames. In such rare and short-term cases, use of longer term temporal information hurts rather than helps classification/detection performance. [5] shows that long short-term memory (LSTM) leads to false detections on an empty (all black) video frame when used for vocal fold segmentation despite use of high-speed endoscopy video.

4.3 Method

The goal of the proposed LARNet and LARNet-STC are automated detection of LAR events in laryngoscopy videos. The LARNet is a two-stream deep classification network. The LARNet-STC is built on top of LARNet, which contains two main parts: (1) VFs state

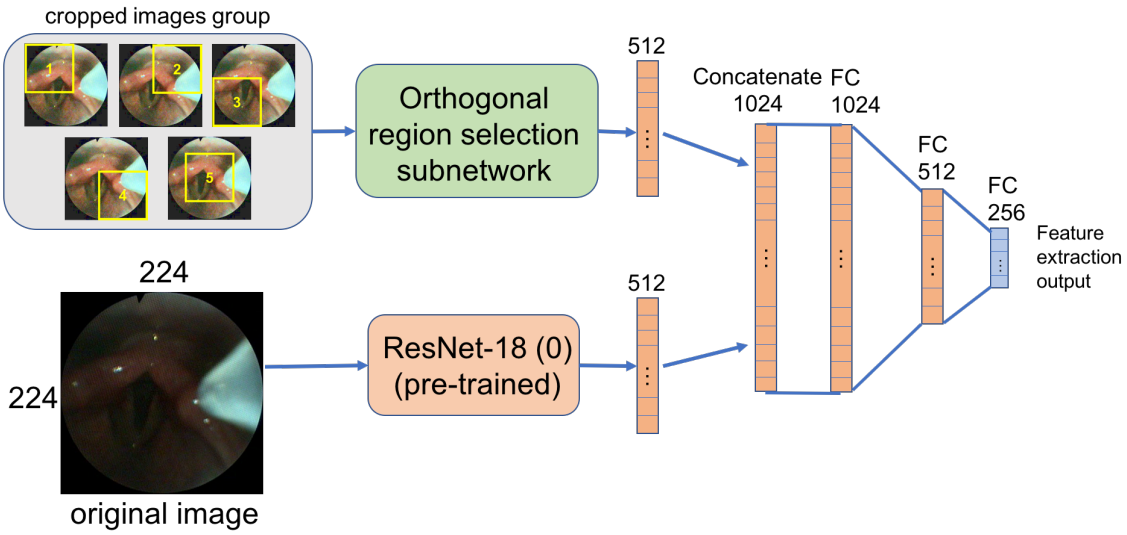


Figure 4.3. Architecture of the proposed VFs state estimation network.

estimation network for single image features extraction, and (2) context-based network for classification.

4.3.1 VFs State Estimation Network (LARNet)

For extracting features, we have first developed a custom, two-stream deep learning network for estimation of VFs open/closed states for each video frame, we call it "LARNet" (Figure 4.3). The two inputs of the network are: (1) original video frame resized to 224×224 pixels; and (2) a set of five cropped sub-regions extracted from the histogram equalized original image. The first input, the original video frame, provides global information from the camera field of view. This original image is fed into an ImageNet [125] pre-trained ResNet-18 convolutional neural network [126] for feature extraction. This first stream produces a 1×512 feature vector. The second input, a set of cropped sub-regions, provides local information. This input is fed into a custom orthogonal region selection subnetwork described in Section 4.3.3. This second stream produces another 1×512 feature vector. The two feature vectors are concatenated to generate a 1×1024 linear vector (concatenate-

1024) followed by a fully-connected layer (FC-1024). Feature extraction is performed by three fully connected layers, FC-1024, FC-512, and FC-256. ReLu (rectified linear unit) activation function and dropout of 0.5 are applied between the concatenate-1024, FC-1024, FC-512, and FC-256 layers. This VF state estimation network produces 1×256 feature vector at the end. In our previous chapter [127], the last fully-connected layer was FC-3 instead of FC-256 for direct classification output.

4.3.2 Image Preprocessing and Subregion Generation

As can be seen in the sample frames shown in Figure 4.2, some of the video frames suffer from very low illumination that heavily affects the visibility of VFs. In the VFs state estimation network, we augment the single-stream network whose input is the original image with a second input and processing stream that deals with a set of cropped images. This second stream aims to address the illumination problems, scale variation, and incorporate local information into the decision process. All original images are first resized to 224×224 pixels (input size of ResNet-18 networks). Histogram equalization [128] is applied to the resized original images in order to improve image contrast and visibility. The VFs cover only a small portion of the endoscope field of view and can appear at different positions in a video frame. To address these issues, five 154×154 subregions (four corners and one center) are cropped from the histogram equalized image as shown in the left side of Figure 4.4. One of these cropped regions is expected to have better coverage of the VFs. These cropped images are then resized to 224×224 pixels and fed into the orthogonal region selection (ORS) subnetwork described below to perform selection and to handle information redundancy.

4.3.3 Orthogonal Region Selection Subnetwork (ORS)

The five cropped and resized images are fed into the ORS subnetwork shown in Figure 4.4. Each cropped image is processed by an ImageNet [125] pre-trained ResNet-18 convolutional neural network [126] for feature extraction. These ResNet-18 networks are

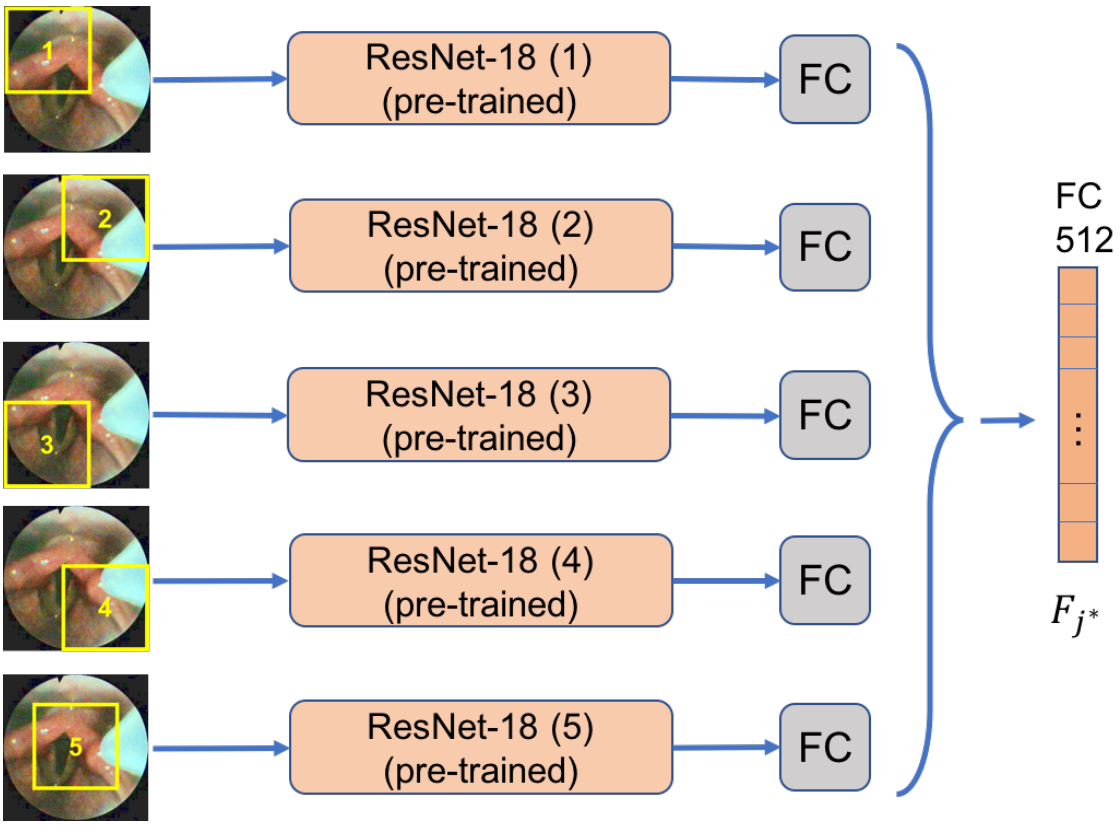


Figure 4.4. Subregion cropping and the Orthogonal Region Selection (ORS) subnetwork. Inputs to the network are five cropped subregions (marked with yellow squares) from the preprocessed image. Output of the network is a 1-D feature vector corresponding to the selected subregion. This vector is selected from F by the index j^* of the minimum value in O. "FC" represents fully-connected layer.

retrained for VFs state estimation. The output of each ResNet-18 network is a 1×512 feature vector corresponding to each cropped region. A 5×512 matrix F is constructed where each row j in F is the feature vector of one subregion. The measure of orthogonality O_j between a subregion R_j and the remaining subregions is defined as:

$$\mathrm{P} = \mathrm{F} \times \mathrm{F}^T - diag\left(\mathrm{F} \times \mathrm{F}^T\right) \tag{4.1}$$

$$O_j = \sum_{k=0}^{4} P_{jk} \tag{4.2}$$

$$j^* = \underset{j}{\operatorname{argmin}}(O_j) \tag{4.3}$$

Each row j in P denotes the dot product between the subregion feature j and others. In equation(4.2), O is a 5-by-1 matrix representing the sum of each row in P as a final measure of orthogonality for each subregion. The 1×512 feature vector F_{j^*} with minimum value in O is selected as the local feature vector used in VFs state estimation.

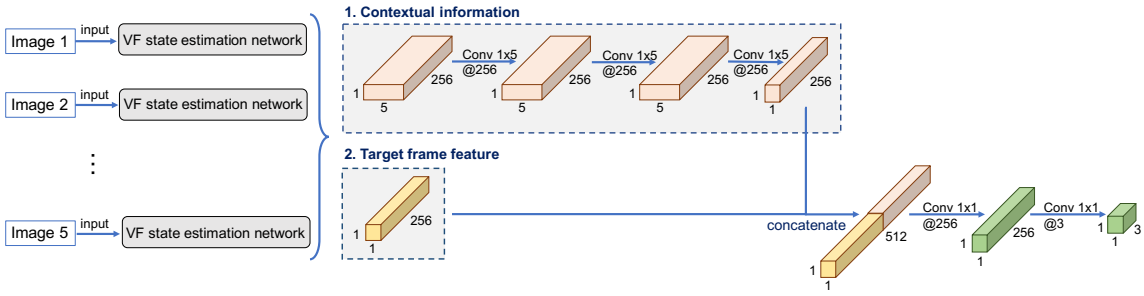


Figure 4.5. Architecture of the proposed spatio-temporal context-based orthogonal region selection network. On top of the VF state estimation networks, a set of fully convolutional layers are inserted to the network to incorporate temporal context. "Conv" represents convolution operation.

4.3.4 Temporal Context-based Orthogonal Region Selection Network of LARNet-STC

Temporal video context is used to improve the state estimation/classification results obtained from independent processing of the video frames by incorporating VF state information from neighboring frames. We propose a simple and efficient deep learning network that combines the spatial information-only VF state estimation network described in Section 4.3.1 with a fully convolutional temporal context feature. The architecture is shown in Figure 4.5. The proposed fully convolutional network for incorporating temporal context features allows an arbitrary number of sequential frames without changing the network architecture. The proposed network concatenates equal sized temporal context and spatial-only VF state estimation feature vectors. The inputs of this network are five sequential images centered at the target frame in the video, we call it an "image block." The target

classification label is the label of the center frame of the image block. For the start and end frames of the video, we pad the video by mirroring the first and last two frames in the video. Then, all five images are preprocessed and fed into the VF state estimation network individually for feature extraction. The VF state estimation network generates 5×256 feature matrix for the five sequential images. According to the manual ground-truth, the average duration of the fully-closed state of the VFs during the LAR events is 5.6 frames. Considering this duration and the trade-off between the hardware memory and execution time, we selected five sequential images as temporal context. We also tested and discussed the proposed network with three and seven sequential images respectively in Section 4.5.2.2.

The 5×256 feature matrix is the spatio-temporal contextual information composed of the five sequential frames centered at the target frame. Then three convolution operations are conducted on the 5×256 feature matrix to further summarize the contextual information. Batch normalization and ReLu activation function are used between each convolution. This stream produces a 1×256 spatio-temporal contextual information vector. The spatio-temporal context vector and the target frame feature vector directly outputted from the VF state estimation network are concatenated to form a 1×512 feature vector. Finally, two additional convolution operations are conducted on the 1×512 feature vector to produce the final classification result. We load the pre-trained weights to the VF state estimation network and freeze them during training. The whole system is end-to-end trainable. Categorical cross-entropy loss and Adam optimizer are used in training. Categorical cross-entropy loss is defined as:

$$CELoss = -\sum_{i}^{C} t_i log\left(f\left(s\right)_i\right) \tag{4.4}$$

$$f\left(s\right)_i = \frac{e^{s_i}}{\sum_{j}^{C} e^{s_j}} \tag{4.5}$$

where C is the total number of classes; t_i, s_i are the ground-truth and network output for the class i respectively; and $f\left(s\right)_i$ is Softmax activation function.

4.3.5 Dealing with Imbalanced Data

LAR is a rare event. Video frames where the VFs are fully closed occur far less frequently than video frames where the VFs are fully or partially open. The closed state only occupies ~7% of the video frames. The occlusion states where the VFs are either occluded or out of the field-of-view occupy ~9% of the video frames. To deal with large data imbalance in VF state estimation, we propose to use an approach that combines random over-sampling (ROS) [59] and random under-sampling (RUS) [60] schemes described in Section 6.2. In each epoch, we randomly generate k image blocks for the context-based orthogonal region selection network to train. The class of the image block is the class of the center frame. First, we randomly select $k/3$ image blocks from all training non-LAR image blocks, and augment with random contrast, zoom in, rotation, and flip image operations. Next, we randomly select $k/6$ image blocks from all training LAR image blocks, then randomly select $k/6$ image blocks from all training occlusion image blocks. Both selected LAR and occlusion image blocks are transformed with random contrast, zoom-in, rotation, and flip to double the training set. Since the non-LAR class is easy to train as they have sufficient data, we chose to raise the portion of LAR and occlusion image blocks in the training data. All selected non-LAR, LAR, and occlusion image blocks are merged together, shuffled, and used for network training. In our experiment, we heuristically pick $k = 300$.

4.4 Experimental Results of LARNet

Our dataset consists of 58 labeled human laryngoscopy videos collected from twenty participants with a 3.7-mm outer-diameter endoscope (11302BD2, Karl Storz), at a frame rate of 30 FPS (frames per second). We randomly separated these videos into training videos and testing videos, resulting in 6,828 frames in training dataset and 2,274 frames in testing dataset. Frame size for all the videos are $480 \times 720 \times 3$. The training dataset contains 5,840 non-LAR frames, 394 LAR frames, and 594 occlusion frames. The test dataset

contains 1,837 non-LAR frames, 251 LAR frames, and 186 occlusion frames. Frames are labeled as "occlusion" if more than half of the VFs are occluded, disappear, or couldn't be recognized in the image. "LAR" class includes fully closed and partially closed VF frames. Partially closed state happens before and after the fully closed state, usually within one or two time steps, and the remained partially open VFs area only occupies within 10% of the full open VFs area. All remaining frames are labeled as "non-LAR". The training and testing procedures are ran on a single GTX 1060 6GB GPU. Testing speed is 27 FPS. System performance is evaluated in terms of precision, recall, and F1-score:

$$precision = \frac{TP}{TP + FP} \tag{4.6}$$

$$recall = \frac{TP}{TP + FN} \tag{4.7}$$

$$F1score = \frac{2 * precision * recall}{precision + recall} \tag{4.8}$$

where TP represents truth positive, FP represents false positive, and FN represents false negative outputs.

4.4.1 Ablation Study of LARNet

In order to better understand the contributions of the different components of the proposed approach, we have designed and trained three additional networks. The first two networks, LarNet1o and LarNet1e, are single-stream networks with ResNet-34 [129] feature extraction modules trained with original and histogram equalized images respectively. The third network, LarNet2, is a two-stream network similar to the proposed network, but without the ORS subnetwork. LarNet2 has two inputs, one is 224-by-224 original image, the other is 224-by-224 histogram equalized image without cropping. The two inputs are processed by two individual ResNet-18 networks. Output feature vectors from the two ResNet-18 networks are concatenated together, followed by another two fully connected layers like in the proposed network.

Table 4.2. VF state estimation performance.

Methods	Precision			Recall			F1 score		
	non-LAR	LAR	occlusion	non-LAR	LAR	occlusion	non-LAR	LAR	occlusion
LarNet1o: Single-stream, ResNet-34 original image	**0.9989**	0.5355	**0.9999**	0.9497	0.9957	0.4585	0.9737	0.6964	0.6288
LarNet1e: Single-stream, ResNet-34 histogram equalized image	0.9972	0.7214	0.9651	0.9676	0.9915	0.8098	0.9822	0.8351	0.8806
LarNet2: Two-stream, ResNet-18 original + histogram equalized image	0.9977	0.6057	0.9545	0.9427	**0.9999**	0.7171	0.9694	0.7544	0.8189
PROPOSED: Two-stream, ResNet-18 original + histogram equalized image Orthogonal Region Selection Subnetwork	0.9967	**0.8007**	**0.9999**	**0.9849**	0.9915	**0.8341**	**0.9908**	**0.8859**	**0.9096**

4.4.2 Results of LARNet

A summary of our VF state estimation performance analysis results are reported in Table 4.2. As we can see from the first two rows of Table 4.2, LarNet1e out-performs LarNet1o. Both networks have identical single-stream architectures using ResNet-34 feature vectors. These results indicate that image preprocessing, specifically histogram equalization in this case, improves the classification accuracy. In the third row of Table 4.2, LarNet2, the two-stream network, which simply concatenates features from the original image and histogram equalized images, respectively, results in the highest recall of LAR. However, without the ORS subnetwork as in the proposed system, this custom network doesn't perform as well as LarNet1e using the histogram equalized images.

The results for the proposed method that combines image preprocessing, two-stream processing, and an ORS subnetwork are shown in the last row of Table 4.2. Precision for the non-LAR cases is slightly lower compared to LarNet1o ($\leq 0.22\%$). Recall for the LAR cases is slightly lower than LarNet2 ($\leq 0.84\%$). However, in all other aspects, the proposed network outperforms the LarNet1o, LarNet1e, and LarNet2 networks. LAR detection precision improves by 26.52% compared to LarNet1o, and 7.93% compared to LarNet1e. As a result, the proposed network improves at least 5.08% in F1 score of LAR, 2.90% in F1 score of occlusion, and 0.86% in F1 score of non-LAR. These results demonstrate that the use of image preprocessing, fusion of global and local information through two-stream network, and the proposed orthogonal region selection subnetwork improve the VF state estimation process. The low occlusion recall values from all four networks are due to two main factors: labeling protocol and LAR-occlusion appearance similarity. Independent of VF open/close state, our ground-truth labeling protocol marks all cases where more than half of the VFs are not visible as occluded. The glottal region, the gap between the VFs, is not visible when either the VFs are closed or occluded, resulting in confusion between occlusion and LAR classes.

4.5 Experimental Results of LARNet-STC

Our dataset consists of 58 labeled human laryngoscopy videos collected from twenty participants with a 3.7-mm outer-diameter endoscope (11302BD2, Karl Storz), at a frame rate of 30 fps (frames per second). The study was approved by the Institutional Review Board of the University of Missouri. We randomly separated all videos into training videos and testing videos, resulting in 9,102 frames in total, of which 6,828 frames (75%) are used as training data and the other 2,274 frames (25%) are used as testing data. Frame size for all the videos are $480 \times 720 \times 3$. The training dataset contains 5,840 non-LAR frames, 394 LAR frames, and 594 occlusion frames. The test dataset contains 1,837 non-LAR frames, 251 LAR frames, and 186 occlusion frames. Frames are labeled as "occlusion" if more than half of the VFs are occluded, disappear, or couldn't be recognized in the image. "LAR" class includes fully closed and partially closed VF frames. A partially closed state happens before and after the fully closed state, usually within one or two time steps, and the remained partially open area only occupies within 10% of the full open VFs area. All remaining frames are labeled as "non-LAR." Directly classifying video frames into VFs' open, closed, or occluded states is a challenging task because of numerous factors such as anatomical variations; diverse illumination conditions, artifacts, and camera focus problems, and motion blur; scale variations; camera or patient motion; partial or full visual obstruction, saliva accumulation, camera position, and motion of the surrounding anatomical structure. A summary of the challenging cases and their frequency in the training and test datasets is summarized in Table 4.1. A single video frame can suffer from multiple types of problems at the same time (i.e. an out-of-focus image with low-light conditions). In those cases, the same frame is counted towards all the associated problem cases. The training and testing procedures are run on a single GTX 1060 6GB GPU.

System performance is evaluated in terms of precision, recall, and F1-score:

$$precision = \frac{TP}{TP + FP} \tag{4.9}$$

Table 4.3. Contribution of the different network components to the classification performance.

Network	Backbone	# Streams	Hist Equal	ORS	Temporal Context	Average F1 score
LARNet-S1R	ResNet-34	1	N	N	N	0.7663
LARNet-S1Re	ResNet-34	1	Y	N	N	0.8993
LARNet	ResNet-18	2	Y	Y	N	0.9288
LARNet-STC	LARNet	2	Y	Y	Y	0.9402

$$recall = \frac{TP}{TP + FN} \tag{4.10}$$

$$F1score = \frac{2 * precision * recall}{precision + recall} \tag{4.11}$$

where TP represents true positive, FP represents false positive, and FN represents false negative outputs.

4.5.1 Ablation Study of LARNet-STC

The contribution of the different network components to the classification performance is summarized in Table 4.3. As we can see in the first two rows of Table 4.3, under the same classification network, the histogram equalization process helps improve the average F1 score by 13.30%. The second, third, and fourth rows demonstrate that incorporating the ORS subnetwork increases the average F1 score by 2.95% and temporal context information increases the average F1 score by 1.14%.

In order to better understand the contributions of the different components of the proposed approach, we have designed and trained five additional networks for spatial only single image classification, and four additional networks for spatio-temporal context-based image classification whose features are summarized in Table 4.4. For reference, Table 4.5 provides the corresponding number of parameters, memory cost, and the computational cost of each of those networks. For the single image classification, we compare

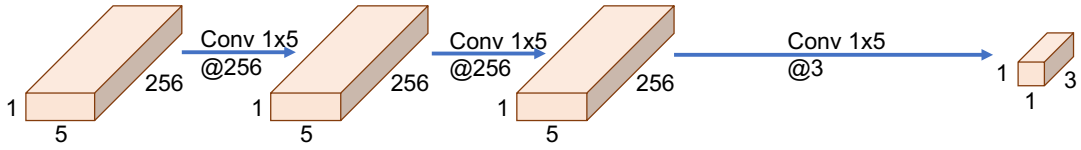
Table 4.4. Feature summary of the proposed network and the comparison networks. The architectural differences between the four spatio-temporal context-based networks are illustrated in Figure 4.6.

Network	Backbone	# Streams	Hist Equal	ORS	Temporal Context
LARNet-S1V	VGG-13	1	N	N	N
LARNet-S1Ve	VGG-13	1	Y	N	N
LARNet-S1R	ResNet-34	1	N	N	N
LARNet-S1Re	ResNet-34	1	Y	N	N
LARNet-S2Re	ResNet-18	2	Y	N	N
LARNet	ResNet-18	2	Y	Y	N
LARNet-ST2Le-Conv	LARNet	2	Y	Y	Y
LARNet-ST2Le-Elem	LARNet	2	Y	Y	Y
LARNet-ST2Le-Elem-Flip	LARNet	2	Y	Y	Y
LARNet-ST2Le-LSTM	LARNet	2	Y	Y	Y
LARNet-STC	LARNet	2	Y	Y	Y

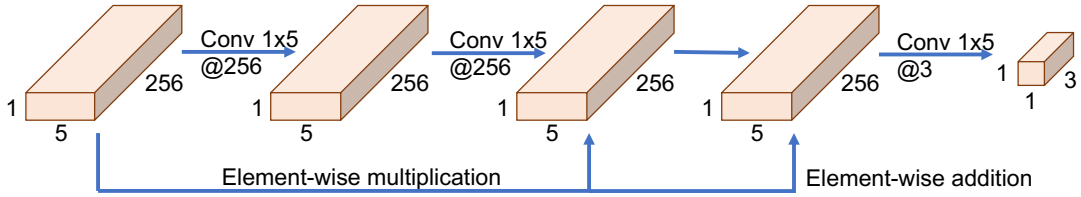
Table 4.5. Networks Features

Network	Number of parameters (million)	Memory cost (Megabyte)	Computational cost (frames per second)
LARNet-S1V	128.96	515.85	84.02
LARNet-S1Ve	128.96	515.85	84.02
LARNet-S1R	24.45	97.89	125.08
LARNet-S1Re	24.45	97.89	125.08
LARNet-S2Re	23.40	93.69	111.57
LARNet	24.45	97.89	27.01
LARNet-ST2Le-Conv	71.00	284.24	4.84
LARNet-ST2Le-Elem	71.00	284.24	4.83
LARNet-ST2Le-Elem-Flip	71.00	284.24	4.82
LARNet-ST2Le-LSTM	71.21	285.08	5.09
LARNet-STC	71.46	286.07	4.27

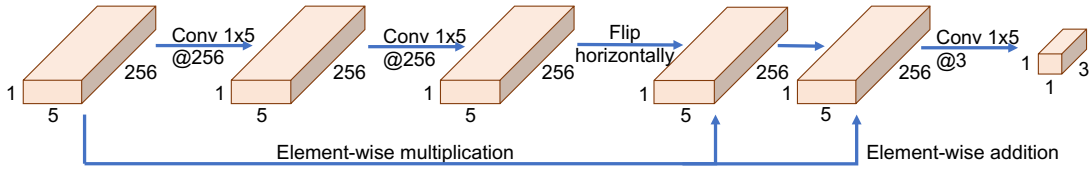
our proposed VF state estimation network with the other five networks. The first two networks, LARNet-S1V and LARNet-S1Ve, are single-stream networks with pre-trained VGG-13 [9]. The second two networks, LARNet-S1R and LARNet-S1Re, are another single-stream networks with pre-trained ResNet-34 [126]. These four networks are without temporal contextual information and trained with original and histogram equalized images respectively. The fifth network, LARNet-S2Re, is a two-stream network similar to the VF state estimation network, but without the ORS subnetwork. LARNet-S2Re has two inputs, one is a 224-by-224 original image, the other is a 224-by-224 histogram equalized image without cropping. The two inputs are processed by two individual pre-trained ResNet-18 networks. Output feature vectors from the two ResNet-18 networks are concatenated together, followed by another two fully connected layers like in the proposed LARNet. The detailed evaluations of these five architectures are reported in rows 1 to 5 of Table 4.6. For the spatio-temporal context-based image classification, we examine four architectures, LARNet-ST2Le-Conv, LARNet-ST2Le-Elem, LARNet-ST2Le-Elem-Flip, and LARNet-ST2Le-LSTM, to abstract and combine the spatio-temporal contextual information on top of the VF state estimation network. The four architectures are shown in Figure 4.6. The inputs of these four architectures are the 5×256 feature matrix outputted directly from the VFs state estimation network for an image block. In the first three architectures, the target image feature and its contextual information inside the 5×256 feature matrix are convolved together without concatenation to get the final classification result. In the fourth architecture, the 5×256 feature matrix is sent into the Long Short-Term Memory (LSTM) to incorporate temporal information, followed by two fully-connected layers for classification. Evaluations of these four architectures are shown in rows 7 to 10 of Table 4.6.



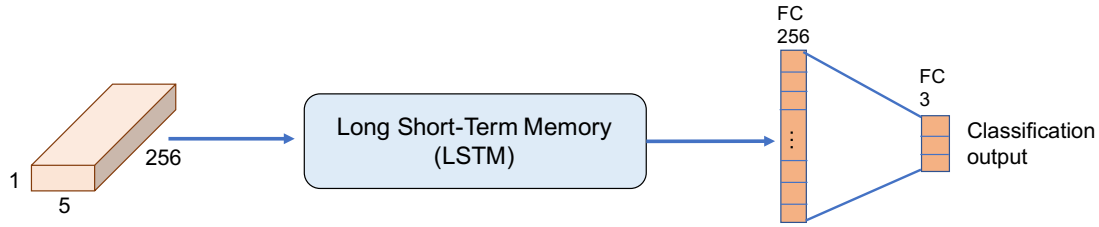
(a) LARNet-ST2Le-Conv: three convolutions on the 5×256 feature matrix



(b) LARNet-ST2Le-Elem: convolutions on the 5×256 feature matrix with element-wise operations



(c) LARNet-ST2Le-Elem-Flip: five convolutions on the 5×256 feature matrix with element-wise operations and horizontally flip



(d) LARNet-ST2Le-LSTM: process the 5×256 feature matrix with Long Short-Term Memory (LSTM)

Figure 4.6. Four different architectures of spatio-temporal context-based networks.

Table 4.6. VFs state estimation performance for the proposed and comparison networks. Features of the compared networks are given in Table 4.4.

Methods	Number of sequential frames	Precision				Recall				F1 score			
		non-LAR	LAR	occlusion	Avg.	non-LAR	LAR	occlusion	Avg.	non-LAR	LAR	occlusion	Avg.
LARNet-S1V	1	0.9200	0.4361	0.2093	0.5218	0.6167	0.5882	0.7463	0.6504	0.7384	0.5009	0.3269	0.5221
LARNet-S1Ve	1	0.9854	0.3094	0.6571	0.6506	0.7672	0.7437	0.8976	0.8028	0.8627	0.4370	0.7588	0.6862
LARNet-S1R	1	**0.9989**	0.5355	**0.9999**	0.8448	0.9497	0.9957	0.4585	0.8013	0.9737	0.6964	0.6288	0.7663
LARNet-S1Re	1	0.9972	0.7214	0.9651	0.8946	0.9676	0.9915	0.8098	0.9230	0.9822	0.8351	0.8806	0.8993
LARNet-S2Re	1	0.9977	0.6057	0.9545	0.8526	0.9427	**0.9999**	0.7171	0.8866	0.9694	0.7544	0.8189	0.8476
LARNet	1	0.9967	0.8007	**0.9999**	0.9324	0.9849	0.9915	0.8341	0.9368	0.9908	0.8859	0.9096	0.9288
LARNet-ST2Le-Conv	5	0.9956	0.8339	0.9532	0.9276	0.9831	0.9602	0.8763	0.9399	0.9893	0.8926	0.9132	0.9317
LARNet-ST2Le-Elem	5	0.9978	0.8356	0.9647	0.9327	0.9842	0.9721	**0.8817**	0.9460	0.9910	0.8987	**0.9213**	0.9370
LARNet-ST2Le-Elem-Flip	5	0.9978	0.8316	0.9759	0.9351	0.9837	0.9841	0.8710	**0.9463**	0.9907	0.9015	0.9205	0.9376
LARNet-ST2Le-LSTM	5	0.9940	0.8445	0.9697	0.9361	0.9880	0.9522	0.8602	0.9335	0.9910	0.8951	0.9117	0.9326
PROPOSED: LARNet-STC	5	0.9945	**0.8686**	0.9647	**0.9426**	**0.9907**	0.9482	**0.8817**	0.9402	**0.9926**	**0.9067**	**0.9213**	**0.9402**
PROPOSED: LARNet-STC	3	0.9934	0.8345	0.9432	0.9237	0.9809	0.9442	0.8925	0.9392	0.9871	0.8860	0.9171	0.9301
PROPOSED: LARNet-STC	7	0.9939	0.8191	0.9583	0.9238	0.9809	0.9562	0.8656	0.9342	0.9874	0.8824	0.9096	0.9265

Table 4.7. Five-fold cross-validation results for the proposed LARNet-STC network.

K-fold	Number of videos	Precision				Recall				F1 score			
		non-LAR	LAR	occlusion	Avg.	non-LAR	LAR	occlusion	Avg.	non-LAR	LAR	occlusion	Avg.
Fold 1	12	0.9999	0.8189	0.9320	0.9169	0.9787	0.9999	0.9999	0.9928	0.9892	0.9004	0.9648	0.9515
Fold 2	12	0.9988	0.7673	0.9632	0.9098	0.9865	0.9810	0.8239	0.9305	0.9926	0.8611	0.8881	0.9139
Fold 3	12	0.9999	0.7986	0.9713	0.9233	0.9763	0.9999	0.9916	0.9893	0.9880	0.8880	0.9814	0.9525
Fold 4	11	0.9973	0.8112	0.8478	0.8854	0.9739	0.9915	0.9512	0.9722	0.9855	0.8923	0.8966	0.9248
Fold 5	11	0.9884	0.9059	0.7639	0.8861	0.9653	0.9809	0.8730	0.9397	0.9767	0.9419	0.8148	0.9111
Average	11.6	0.9969	0.8204	0.8956	0.9043	0.9761	0.9906	0.9279	0.9649	0.9864	0.8967	0.9091	0.9308

4.5.2 Results of LARNet-STC

A summary of our VFs state estimation performance analysis results based on the testing set is reported in Table 4.6. Five-fold cross-validation results for the proposed LARNet-STC network are listed in Table 4.7.

4.5.2.1 Single Image Classification using LARNet

In this subsection, we will discuss single image classification networks without temporal contextual information. As we can see from the first four rows of Table 4.6, LARNet-S1Ve out-performs LARNet-S1V, and LARNet-S1Re out-performs LARNet-S1R. LARNet-S1Ve and LARNet-S1V have identical VGG-13 single-stream architectures, LARNet-S1Re and LARNet-S1R have identical ResNet-34 single-stream architectures. These results (the first four rows of Table 4.6) indicate that, under the same classification network, image preprocessing, specifically histogram equalization, in this case, improves the classification accuracy.

In the fifth row of Table 4.6, LARNet-S2Re, the two-stream network, which simply concatenates ResNet-18 features from the original image and histogram equalized images, respectively, results in the highest recall of LAR. However, without the ORS subnetwork as in the proposed system, this custom network doesn't perform as well as LARNet-S1Re using the histogram equalized images. The results for the proposed VFs state estimation network that combines image preprocessing, two-stream processing, and an ORS subnetwork are shown in the sixth row of Table 4.6. Precision for the non-LAR cases is slightly lower compared to LARNet-S1R ($\leq 0.22\%$). Recall for the LAR cases is slightly lower than LARNet-S2Re ($\leq 0.84\%$). However, in all other aspects, the VFs state estimation network outperforms the LARNet-S1V, LARNet-S1Ve, LARNet-S1R, LARNet-S1Re, and LARNet-S2Re networks. LAR detection precision improves by 49.13% compared to LARNet-S1Ve, 26.52% compared to LARNet-S1R, and 7.93% compared to LARNet-S1Re. As a result, the proposed network improves at least 5.08% in the F1 score of LAR,

2.90% in the F1 score of occlusion, and 0.86% in the F1 score of non-LAR. The above results demonstrate that the use of image preprocessing, the fusion of global and local information through the two-stream network, and the proposed orthogonal region selection subnetwork improve the VFs state estimation process without the knowledge of contextual information.

4.5.2.2 Spatio-temporal Context-based Image Classification using LARNet-STC

In order to improve single image classification results, we combined the proposed VF state estimation network with spatio-temporal video contextual information. The results of spatio-temporal context-based networks are shown in the last five rows in the Table 4.6. As we can see in the table, combining temporal information into the classification networks brings over 3.09% improvements in the precision of LAR and over 2.61% improvements in the recall of occlusion. Except for the F1 score of non-LAR in LARNet-ST2Le-Conv, all the spatio-temporal context-based networks achieve overall improvements in the F1 score of all three VF states compared to the single image classification networks. The results for the proposed LARNet-STC shown in the 11th row of Table 4.6 out-perform the other four context-based networks, LARNet-ST2Le-Conv, LARNet-ST2Le-Elem, LARNet-ST2Le-Elem-Flip, and LARNet-ST2Le-LSTM. The precision of LAR in the 11th row of the table is at least 2.41% better than the other four context-based networks and 6.79% better than the VFs state estimation network. The results of the proposed network demonstrate that the concatenation of the spatio-temporal contextual information vector and the target image feature vector works more efficiently than convolution only on the spatio-temporal contextual information vector alone as well as LSTM, indicating the importance of considering jointly single-frame-based classification and context-based classification. The proposed spatio-temporal context-based orthogonal region selection network improves 0.18%, 2.08%, and 1.17% of the F1 score of non-LAR, LAR, and occlusion respectively compared to the VF state estimation network of our previous work.

To evaluate the effect of temporal window on performance, we tested the proposed LARNet-STC network with temporal windows of three and seven frames (Table 4.6 last two rows). Temporal windows of three frames improves average recall and F1 score compared to the spatial-only LARNet, but not as much as the proposed LARNet-STC network with five frames. The proposed LARNet-STC network with longer temporal windows leads to lower average precision, recall, and F1 score compared to the spatial-only LARNet. Best performance is observed for temporal window size of five frames. This is the window size closest to average duration of fully-closed state of the VFs during LAR events (5.6 frames).

4.5.2.3 K-fold Cross-validation of LARNet-STC

We have further evaluated the performance of the proposed LARNet-STC network using five-fold cross-validation. The total number of videos in each fold and cross-validation results are reported in the Table 4.7.

In the fifth fold, we notice that the precision of the occlusion class is relatively low. This is because of the manually labeled non-LAR frames predicted as occlusion frames by the network. Further visual inspection of these ”misclassified” frames revealed that in these frames vocal folds were in fact partially occluded by the surrounding tissue and the images were out-of-focus. However, since our occlusion criteria for manual labeling was occlusion of at least half of the vocal folds, these frames were labeled as non-LAR even though they were partially occluded. For these misclassified frames, the second-largest prediction probability happens at the non-LAR class, which means that the system potentially considers these frames as non-LAR, but because of the out-of-focus image conditions and partial occlusion, the system classifies them as occlusion.

For comparison, the same five-fold cross-validation is applied to LARNet and LARNet-ST2Le-Conv, the results are shown in Table 4.8.

4.5.2.4 Statistical Significance of LARNet-STC

We have used the K-fold cross-validation paired t test [130] to evaluate the statistical significance of the results obtained by the proposed networks with respect to each other. The five-fold cross-validated paired t test is defined as:

$$t = \frac{\bar{p} \cdot \sqrt{n}}{\sqrt{\frac{\sum_{i=1}^{n-1}\left(p^{(i)} - \bar{p}\right)^2}{n-1}}} \tag{4.12}$$

where $\bar{p}$ is the mean score difference between the two networks over the five folds, $p^{(i)}$ is the score difference of the i^{th} fold. n is the number of folds, which is five here. The paired t test score and the corresponding p-value between different networks are shown in Table 4.9. We have performed a two-tailed paired t test and computed p-value for a significance level of $alpha = 0.05$. These results show that while the difference between our two spatio-temporal networks (LARNet-ST2Le-Conv and LARNet-STC) is not significant, the difference between spatial-only network LARNet and the spatio-temporal networks LARNet-ST2Le-Conv and LARNet-STC are statistically significant with p-values of 0.029 and 0.004 respectively. These results demonstrate the significance of temporal context in LAR event detection.

Boxplot of the $F1$ scores of the five-fold cross-validation of the LARNet, LARNet-ST2Le-Conv, and LARNet-STC networks is shown in Figure 4.7. The green triangle is the mean over the five folds. As we can see in Figure 4.7, the proposed LARNet-STC shows the best performance with a mean $F1$ score of $\overline{F1} = 0.9308 \pm 0.0179$ (the second

Table 4.8. The five-fold cross-validation results comparison of LARNet, LARNet-ST2Le-Conv, and the proposed LARNet-STC. The first value is mean, the second value is standard deviation.

Network	Mean Precision	Mean Recall	Mean F1
LARNet	0.8546±0.0436	0.9173±0.0226	0.8811±0.0242
LARNet-ST2Le-Conv	0.8913±0.0393	**0.9674±0.0219**	0.9230±0.0282
LARNet-STC	**0.9043±0.0157**	0.9649±0.0255	**0.9308±0.0179**

Table 4.9. Statistical significance analysis. The five-fold cross-validated paired t test of the F1 scores of three different networks.

Methods	t-score	p-value (alpha=0.05)
LARNet-ST2Le-Conv & LARNet-STC	-1.0120	0.3688
LARNet-ST2Le-Conv & LARNet	3.3216	0.0293
LARNet-STC & LARNet	5.9344	0.0040

value is the standard deviation) over the five folds, which shows a nearly 5% improvement after introducing the temporal context compared to the LARNet ($\overline{F1} = 0.8811 \pm 0.0242$). The medians of the $F1$ scores over the five folds of LARNet-ST2Le-Conv and LARNet-STC are similar. However, the mean $F1$ score over the five folds of LARNet-ST2Le-Conv ($\overline{F1} = 0.9230 \pm 0.0282$) is lower than the LARNet-STC. The 25th percentile of the LARNet-ST2Le-Conv is slightly higher than the one of the LARNet-STC ($\geq$0.22%). However, the 75th percentile of the LARNet-STC is higher than the one of the LARNet-ST2Le-Conv ($\geq$0.5%). As we can see in the boxplot, the range of the F1 score of the proposed LARNet-STC is smaller than the other two networks, indicating that the proposed LARNet-STC is more robust across all five folds compared to the others. In the boxplot, both of the spatio-temporal networks LARNet-ST2Le-Conv and LARNet-STC improve the mean F1 score by 4.20% and 4.97% respectively compared to the spatial-only LARNet, demonstrating that the temporal contextual information further improves the accuracy compared to the spatial-only information.

4.5.2.5 Quantitative Evaluation of LAR Event Durations

We have further evaluated the proposed LARNet-STC based on the event durations. For every video in the five-folds cross-validation output, we computed the frame error e between the ground truth number of frames and the predicted number of frames for every LAR event i, which is defined as:

$$e_i = |f_{GT_i} - f_{pred_i}|, i \in N \tag{4.13}$$

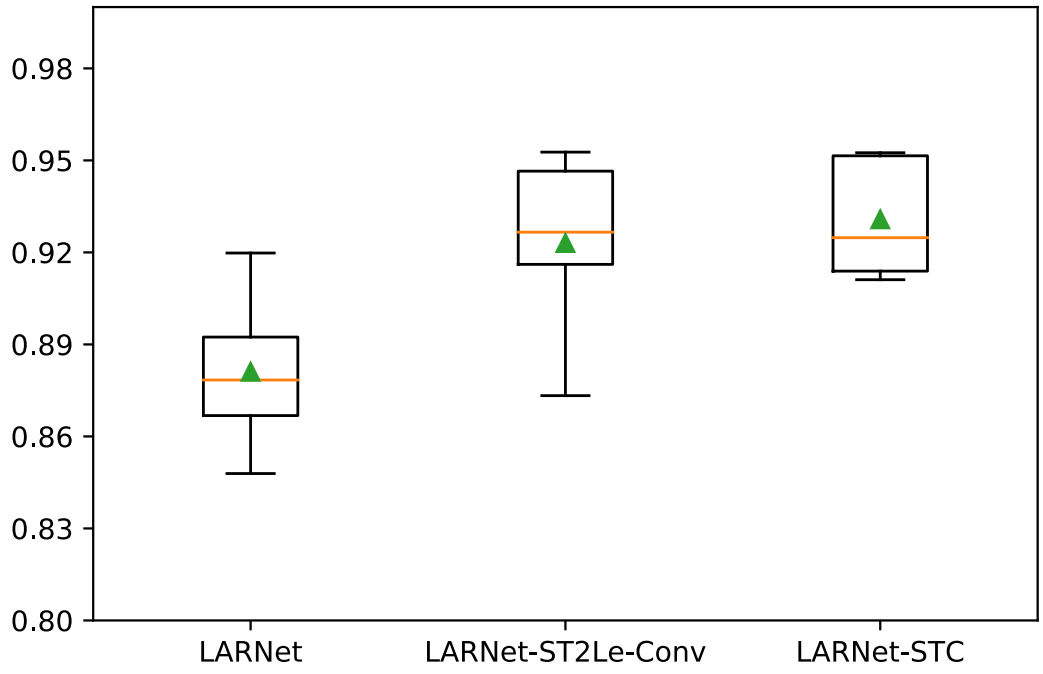


Figure 4.7. Boxplot of the F1 scores for the five-fold cross-validation of three proposed networks. The green triangle is the mean across five folds.

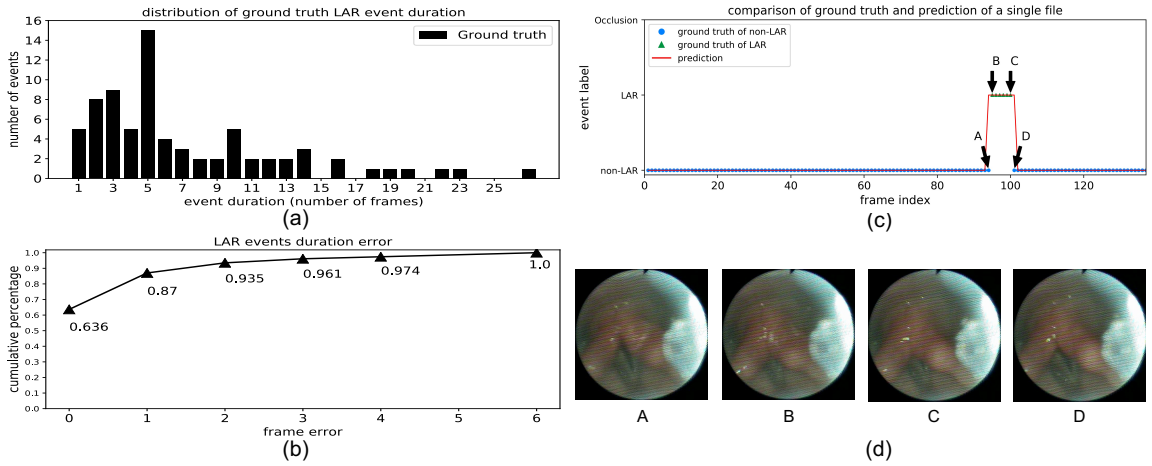


Figure 4.8. Quantification evaluation of LAR event durations (number of frames). (a) Histogram of the distribution of ground truth LAR event durations. (b) cumulative distribution of the frame error of LAR event prediction. (c) Comparison of the ground truth and prediction of VF states for a single video. (d) Sample original video frames at timestamps A, B, C, and D in (c).

where f_{GT_i} is the ground truth number of frames for the event i, f_{pred_i} is the predicted number of frames for the event i. N is total number of LAR events.

Figure 4.8(a) shows the distribution of ground truth LAR event durations. In the ground truth durations of LAR events, 77.33% of the LAR events last within 10 frames (0.33 seconds). Figure 4.8(b) shows the cumulative distribution of frame error n ($\frac{\{\#i|e_i \leq n\}}{N}, i \in N$) for LAR events. As we can see in Figure 4.8(b), 63.6% of the LAR events have no error, and 87% of the LAR events have equal to or less than one frame error, which corresponds to an error rate of 17.86%, considering that the average duration of the fully-closed state of the VFs during the LAR event is 5.6 frames. Most frame errors occur at the beginning and end of the LAR events. This type of error is due to the discrepancy between the ground truth and prediction of the start and end points of a LAR event. As an example, Figure 4.8(c) shows the comparison of ground truth and prediction of VF states for a single video, and Figure 4.8(d) shows the sample original video frames. We pick four timestamps A, B, C, and D from Figure 4.8(c), and display the sample original video frames at these four timestamps in Figure 4.8(d). As we can see in the sample original video frames, the FEEST triggered laryngeal adductor reflex (LAR) starts at timestamp A, leading to the LAR prediction in image A. In images B and C (timestamps B and C), the vocal folds are fully closed and correctly detected as LAR events. Image D is the first frame VFs start to open. Manual ground truth lists it as non-LAR, while the proposed network predicts it as LAR.

4.5.2.6 Using Glottal Region Segmentation for LAR Event Detection

To further justify the need for an explicit LAR/non-LAR classification network as proposed in this chapter, we ran three deep-learning-based glottal region (region between the VFs) segmentation networks on our test laryngoscopy videos, and inferred the LAR versus non-LAR states by thresholding the glottal region area indicated by the resulting segmentation masks. The evaluated segmentation networks are: (1) U-LSTM network [5]; (2) FCRN

[6]; and (3) FCRN+HE+ORS, FCRN [6] network with inputs preprocessed as LARNet and LARNet-STC (histogram equalized and cropped using ORS subnetwork). U-LSTM network [5] was recently proposed to study VF vibrations during speech production using high-speed video (HSV) acquired using rigid transoral (through the mouth) laryngoscopy. FCRN [6] is our group's earlier network developed to study VF motion dynamics through glottal region segmentation on videos acquired using flexible transnasal endoscopy. We have tested and evaluated five different area threshold values (200, 250, 300, 500, and 700 pixels) to infer LAR versus non-LAR states from the segmentation masks. A threshold value higher than zero is used because in some LAR cases, a slight opening is still left between the VFs when they are closed. The LAR/non-LAR classification results inferred from glottal region segmentation masks were compared to the ground truth class labels and compared to the performance of the proposed direct classification network LARNet-STC. The F1 scores for non-LAR and LAR frames of different methods are reported in Table 4.10 and Table 4.11 respectively, and shown in Figure 4.9 and Figure 4.10.

Best segmentation-based LAR detection results are obtained for the FCRN+HE+ORS method with an area threshold of 200. While acceptable results are obtained for non-LAR cases (F1=0.89), all segmentation networks fail during the LAR events (best performance is F1=0.38). Accurate detection of the LAR events is the core of the Flexible Endoscopic Evaluation of Swallowing with Sensory Testing (FEESST) used to assess life-threatening VF dysfunctions affecting breathing and swallowing. The low performance by segmentation-based methods is mainly due to false glottal region detections during LAR events. While it is feasible to accurately segment (or track) the VFs during their regular vibratory states (i.e. during speech or breathing). The suddenness and rarity of the LAR events, large appearance variations of the VFs and the glottal region prevent reliable segmentation or tracking of the VFs during the LAR events and lead to much lower scores compared to the proposed network.

Table 4.10. Segmentation-derived LAR/non-LAR classification results. Average F1 scores for non-LAR frames.

Segmentation Algorithm	threshold= 200 pixels	threshold= 250 pixels	threshold= 300 pixels	threshold= 500 pixels	threshold= 700 pixels
U-LSTM [5]	0.8267	0.8196	0.8072	0.7935	0.7506
FCRN [6]	0.8908	0.6859	0.6790	0.6480	0.6428
FCRN [6] + histogram equalization + ORS	0.8933	0.8724	0.8362	0.6620	0.4746
Proposed: LARNet-STC	0.9926	0.9926	0.9926	0.9926	0.9926

Table 4.11. Segmentation-derived LAR/non-LAR classification results. Average F1 scores for LAR frames.

Segmentation Algorithm	threshold= 200 pixels	threshold= 250 pixels	threshold= 300 pixels	threshold= 500 pixels	threshold= 700 pixels
U-LSTM [5]	0.0699	0.0678	0.1020	0.1055	0.1013
FCRN [6]	0.0000	0.1280	0.1304	0.1239	0.1226
FCRN [6] + histogram equalization + ORS	0.3825	0.3817	0.3645	0.2910	0.2428
Proposed: LARNet-STC	0.9067	0.9067	0.9067	0.9067	0.9067

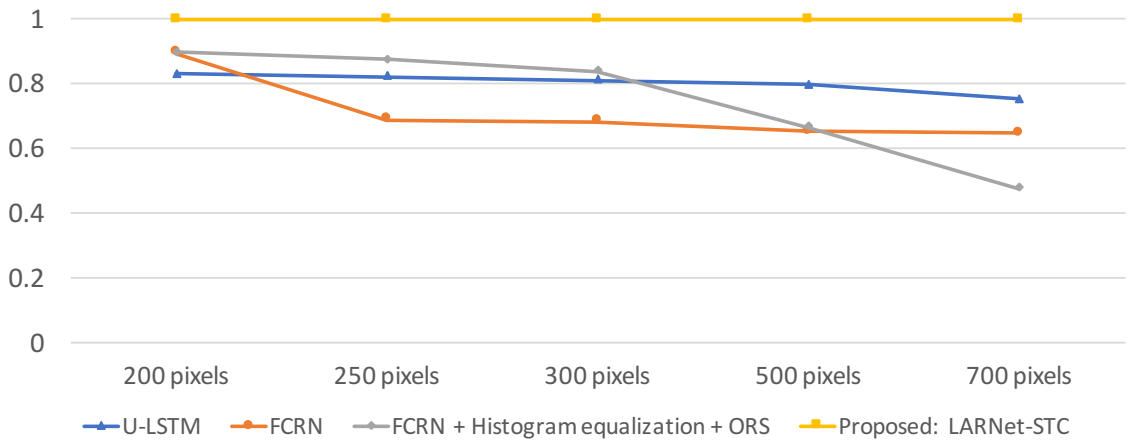


Figure 4.9. Segmentation-derived LAR/non-LAR classification results. Average F1 scores for non-LAR frames. VFs segmentation algorithms (U-LSTM [5], FCRN [6], and FCRN [6] + histogram equalization + ORS) and the proposed LARNet-STC.

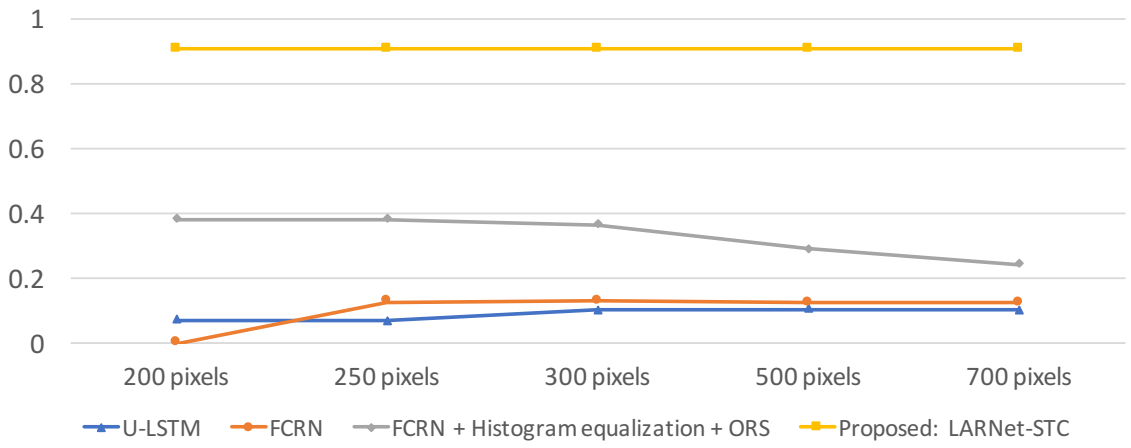


Figure 4.10. Segmentation-derived LAR/non-LAR classification results. Average F1 scores for LAR frames. VFs segmentation algorithms (U-LSTM [5], FCRN [6], and FCRN [6] + histogram equalization + ORS) and the proposed LARNet-STC.

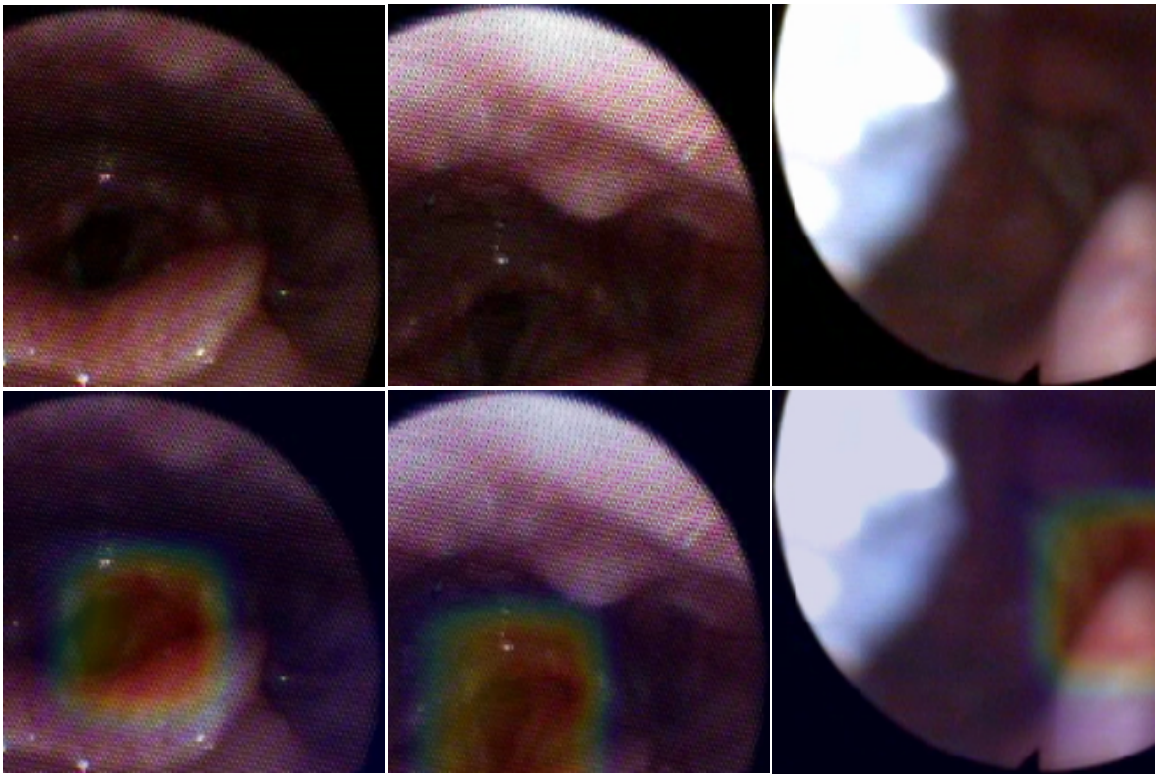


Figure 4.11. Visual explanation of the LARNet-STC network output using Grad-CAM visualization [7]. Top row: subregions automatically selected by Orthogonal Region Selection (ORS) subnetwork. Bottom row: regions corresponding to high score for the predicted class marked with highlights changing from red to blue corresponding to higher to lower impact regions.

4.5.2.7 Visual Explanation for LARNet-STC

Grad-CAM [7] visual analysis software visualizes image regions leading to high scores for the class predicted by a network. We have used Grad-CAM to visually explore the VF state class prediction behavior of the proposed LARNet-STC network. Figure 4.11 demonstrates that: (1) ORS subnetwork successfully selects the image subregions containing the VFs; and (2) the LARNet-STC networks bases its prediction based on VF regions, even though the system does not include an explicit VF segmentation module.

4.5.2.8 Discussion

Figure 4.12 shows the confusion matrix of the results of testing data from the proposed LARNet-STC. As we can see in the confusion matrix, the majority of the errors occur

in predicting some non-LAR images as LAR, LAR as non-LAR, and occlusion as LAR. The errors are due to two main factors: labeling protocol and LAR-occlusion appearance similarity. LAR images include fully closed and partially open/closed images, and the partially open area only occupies within 10% of the fully open VFs area. However, the non-LAR image could have a marginal change in terms of the partially open VFs area compared to the adjacent LAR image and will cause confusion between non-LAR and LAR in the classifier (the middle figure of the first row in Figure 4.13). Independent of VF open/closed state, our ground-truth labeling protocol marks all cases where more than half of the VFs are not visible as occluded. The glottal region, the gap between the VFs, is not visible when either the VFs are closed or occluded, resulting in confusion between occlusion and LAR classes. The middle figure of the third row in Figure 4.13 shows the vocal folds are occluded during the laryngoscopy being pulled out. The third figure of the first row in Figure 4.13 shows the vocal folds are partially occluded (less than a half) by the anatomical structure, but still can be recognized as non-LAR.

The introduction of temporal information for single frame classification only brings limited improvements compared to our previous work. The performance is due to two main reasons: misclassification with the contextual information and the classification error from the previous VFs state estimation network (LARNet). While some of the video frames classification results get corrected with the aid of temporal information, some of the video frames get misclassified. Sample sequential video frames are shown in Figure 4.14. As we can see in the middle figure of Figure 4.14, the ground truth of the frame-8 is occlusion due to the half occluded vocal folds area caused by the artifact, while the prediction of the frame-8 is non-LAR derived by the non-LAR prediction of the frame 6 to frame 10. Randomly separating labeled human laryngoscopy videos into training and testing dataset makes the testing dataset have new imaging problems and anatomical variations that never happened in the training dataset. Under these circumstances, our proposed spatio-temporal

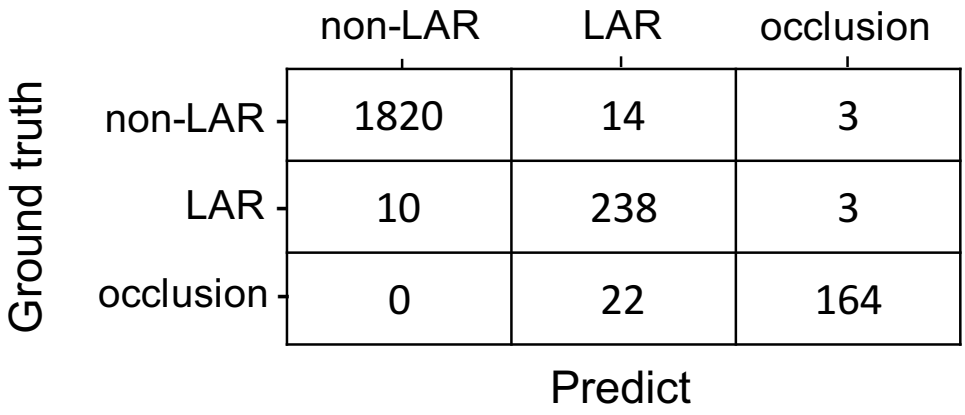


Figure 4.12. The confusion matrix of the results from the proposed context-based orthogonal region selection network (LARNet-STC).

context-based orthogonal region selection network can still achieve over 94% of the testing average F1 score, showing its robustness.

4.6 Conclusion

Laryngeal adductor reflex (LAR) is an airway protection mechanism where the vocal folds (VFs) close abruptly to prevent the entry of foreign materials into the upper airway. VF function and LAR are affected by numerous medical conditions including stroke, tumor, and neurological disorders. LAR impairment increases the risk of aspiration pneumonia and death. In this chapter, we presented deep learning-based systems for automated assessment of the LAR events in laryngeal endoscopy videos. The proposed deep learning networks, LARNet and LARNet-STC, incorporates our novel orthogonal region selection network that acts like an unsupervised spatial attention mechanism and temporal context from the surrounding video frames for robust spatio-temporal reasoning. The network learns to directly map its input (a laryngeal endoscopy image) to a VF open/closed state without prior knowledge of the region of interest (i.e. VF segmentation mask, VF bound-

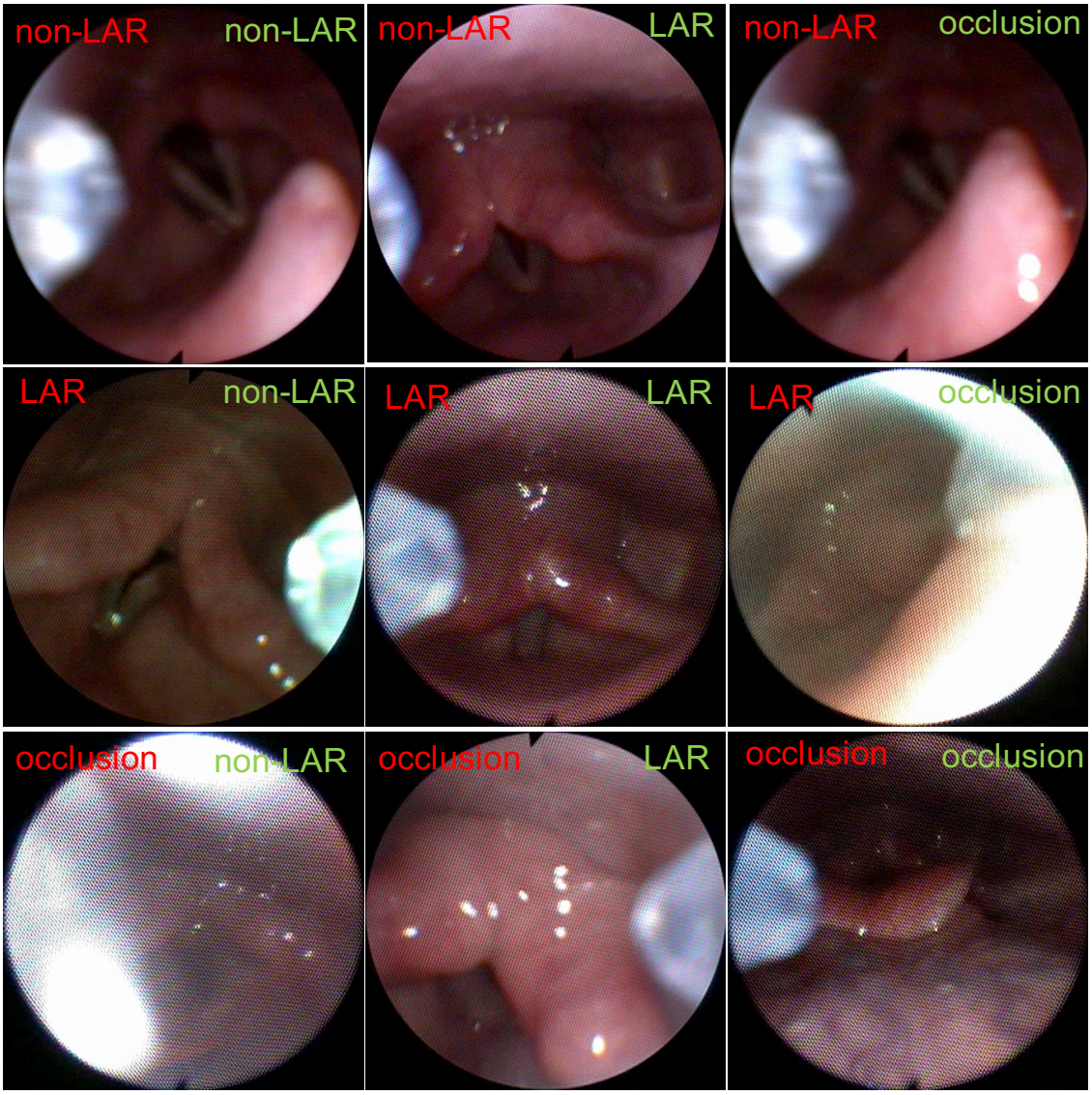


Figure 4.13. Sample outputs from the proposed system. Red label represents ground truth. Green label represents prediction.

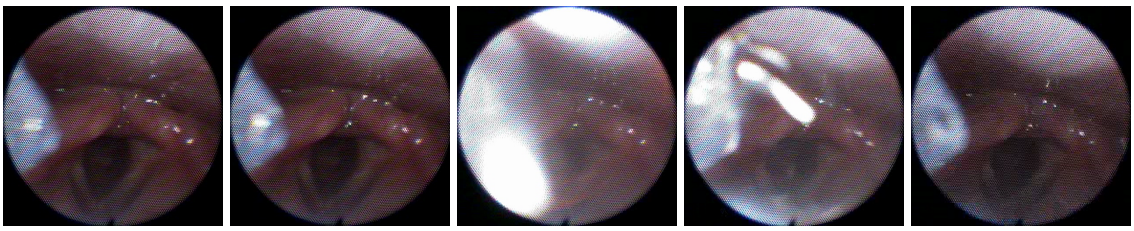


Figure 4.14. Sampled non-LAR sequential video frames (frame 6-10) with visual occlusion from laryngoscopy videos.

ing box, VF motion trajectories). This direct approach considerably reduces the annotation workload needed to train the network. Experimental results showed promising performance towards the automated, objective, quantitative analysis of the LAR events. We have demonstrated that the combination of spatial attention and temporal context greatly improves the LAR detection performance.

We have three main future directions.

(1) Extension of the proposed system to identify different stages of the LAR event: The proposed system classifies its input into open, closed, and occluded states of the VFs. Our first future direction is the extension of the proposed system to identify different stages during the LAR events. Specifically, we are interested in identifying the VF adduction (moving towards) and abduction (parting away) stages. In addition to a more detailed analysis, this extension is expected to further improve LAR duration estimation. While the start of the LAR events is abrupt, the end of the LAR events tends to be more gradually changing to partially open/closed states, which are currently assigned to either open or closed VF states. This extension will resolve ambiguity for these cases where the VFs are in the process of opening or closing and will lead to more precise localization of the LAR start and end time points. Since our current network already involves both spatial and temporal features, drastic network architecture change is not expected. However, the training of the extended network requires considerably more detailed and precise annotations.

(2) Development of a user-friendly full video processing system to ease adoption in clinical settings: We have demonstrated video analytics capabilities on short video clips of interest manually extracted by scientists from clinical test videos. Integration of the current video analytics system into a user-friendly real-time software system that captures and processes the endoscopy video and presents the outputs to the user is needed to facilitate wide clinical adoption. Our first step towards this goal is improving computational cost, and dealing with image quality issues with more robust image preprocessing steps.

(3) Clinical studies to assess the utility of the proposed VF state estimation and LAR detection system: Our interdisciplinary team's ultimate goal is the development of robust machine learning and computer vision systems for objective and quantitative monitoring of disease progression and treatment outcomes associated with VF function and their adoption in clinical settings. The proposed solution will be used as a step forward toward this goal. Our ultimate future direction is to conduct clinical studies to assess the utility of the proposed system in disease vs. healthy behavior discrimination, monitoring disease progression, and treatment efficacy; and to assess its potential for early detection of upper airway dysfunction.

CHAPTER 5

3-D VOLUME SEGMENTATION: ENSEMBLE OF DEEP LEARNING CASCADES WITH GLOBAL SOFT ATTENTION

This chapter introduces an ensemble of deep learning cascade architecture for 2-D image segmentation. Both deep-learning cascades combine unique global soft attention mechanisms, aiming at recall and precision scores respectively. The ensemble mechanism further improves the overall segmentation accuracy of the system. This proposed architecture has been published in [29]. The proposed ensemble of deep learning cascade architecture is applied to 2-D confocal microscopy images for extracting quantitative results for confocal microscopy images.

5.1 Introduction

Detection, segmentation, and quantification of microvascular structures are the main steps towards studying microvascular remodeling. Combined with appropriate staining, confocal microscopy imaging enables exploration of the full 3D anatomical characteristics of microvascular systems. Segmentation of confocal microscopy images is a challenging task due to complexity of anatomical structures, staining and imaging issues, and lack of annotated training data. In this chapter, we propose a deep learning system for robust segmentation of cranial vasculature of mice in confocal microscopy images. The proposed system is an ensemble of two deep-learning cascades consisting of two coarse-to-fine subnetworks with skip connections in between. One cascade aims to improve sensitivity, while the other aims to improve precision of the segmentation results. Our experiments on mice

cranial vasculature showed promising results achieving segmentation accuracy of 92.02% and dice score of 81.45% despite being trained on very limited confocal microscopy data.

Serious intracranial pathologic conditions, such as dural sinus thrombosis, dural arteriovenous fistulas, and aneurysms, involve the vessels, not of the brain itself, but its outer fibrous membrane, the dura mater. These conditions, resulting in significant neurologic morbidity and reduced cognitive abilities [131][132], have been associated with vascular abnormalities. Meningeal vascular networks contribute to brain metabolic clearance and venous blood outflow. They constantly adjust to tissue metabolic demands through structural and functional remodeling. Defective vascular remodeling under certain pathological conditions leads to tissue damage and limits its repair. Acute damage to meningeal vasculature caused by traumatic brain injury, resulting in disruption of meningeal vascular integrity and peripheral immune response can lead to life-threatening situations [132]. While impaired vascular integrity, capillary rarefaction, and aberrant angio-architecture can also develop under chronic conditions, for example, associated with sex hormone deprivation [133].

Detection, segmentation, and quantification of microvascular structures are the main steps towards studying microvascular remodeling. Confocal microscopy [134] allows 3-D image capture using optical sectioning or depth discrimination by blocking light emitted from out-of-focus planes. Each single focus image captures the details of the specimen regions that lie close to its focal plane, while the remaining regions are imaged with poor contrast. Combined with appropriate staining, confocal microscopy imaging enables exploration of the full 3-D anatomical characteristics of microvascular systems.

While segmentation of vessels on traditional angiogram-based imagery or retinal imagery has advanced considerably [135, 136, 22, 83, 84, 85, 137, 138], segmentation of microvasculature in confocal microscopy images remains to be a challenging task. The main challenges are due to complexities of anatomical structures such as irregular shape and varying scale of the vessels; staining issues such as non-homogeneous staining within

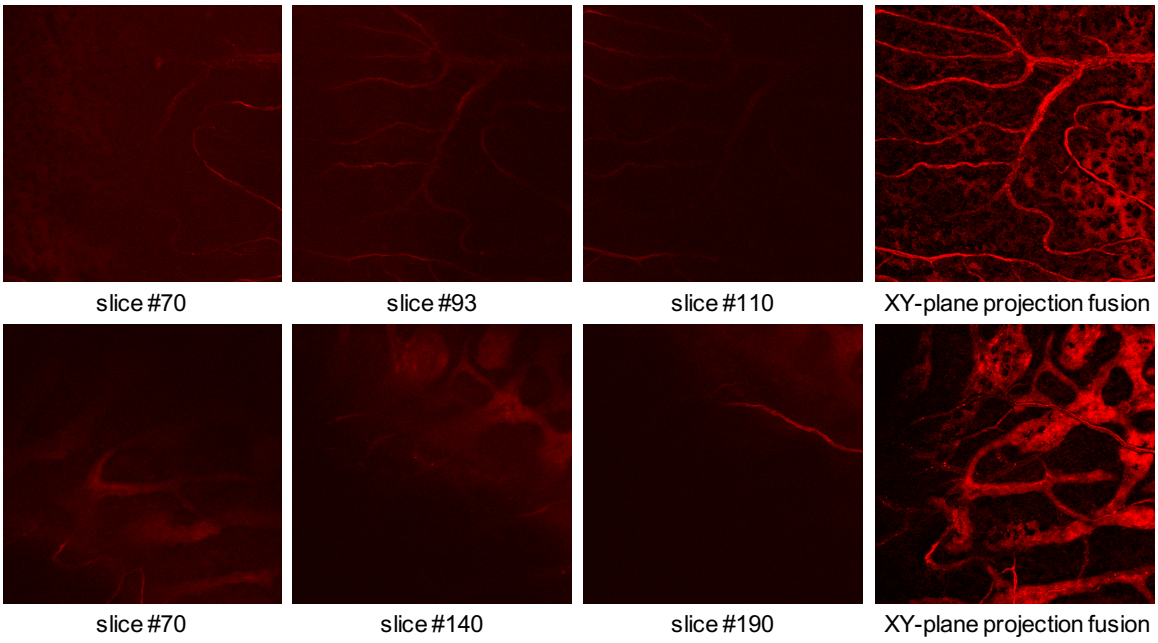


Figure 5.1. Sample blood microvascular structures imaged using confocal microscopy. The first three columns show sample single focus slices. The last column shows fused multi-focus image.

the lumen and excessive stain in the background due to leakage; imaging issues such as low contrast or background clutter due to out of focus structures. Figure 5.1 that shows sample slices and xy-plane projections for two sample confocal microscopy stacks illustrates some of these challenges. In addition to aforementioned challenges, unlike angiogram-based or retinal vessel imagery, manually annotated training data needed for supervised machine learning approaches is severely lacking for confocal microscopy images of microvasculature. This is due to the size of data (hundreds of slices per stack), complexity of the vascular structures, and difficulty of 3-D annotation.

In this chapter, we propose a segmentation system for the confocal microscopy image segmentation, which is a fusion of two deep-learning cascades. The two cascades focus on improving sensitivity (recall) and precision of the segmentation results respectively. To compensate for limited confocal microscopy training data, the proposed network is first

trained with an epifluorescence microscopy image dataset, then fine-tuned with a small set of fused confocal microscopy images of mice cranial microvasculature.

5.2 Ensemble of Deep Learning Cascades for Vessel Segmentation

5.2.1 Image Preprocessing

The proposed preprocessing scheme consists of two steps: (1) multi-focus image fusion, and (2) contrast enhancement. Each confocal microscopy image stack consists of hundreds of single focus slices capturing only a small portion of the microvascular network. We use the multi-focus image fusion approach described in [139] to produce a single multi-focus image out of hundred of single-focus slices within a confocal microscopy stack. The resulting multi-focus image (as the original set of slices) typically suffers from staining issues (i.e. non-homogeneous staining within the lumen and excessive stain in the background due to leakage), imaging issues (i.e. low contrast), and background clutter due to out of focus structures. To improve image contrast, we apply adaptive histogram equalization [140] to the fused multi-focus image. Sample fused, multi-focus, confocal microscopy images before and after adaptive histogram equalization are shown in Figure 5.2.

5.2.2 Network Architecture

For robust and precise segmentation of microvascular structures on confocal microscopy images we have developed two deep learning network cascades: (1) deep binary attention cascade (DBAC), and (2) deep distance map attention cascade (DDMAC). The DBAC network is designed for improving sensitivity (recall) scores, while the DDMAC network is designed for improving precision scores of microvascular image segmentation results. Each cascade generates two outputs: an intermediate output from the first subnetwork acting as attention map for the second network and a final refined segmentation mask from the sec-

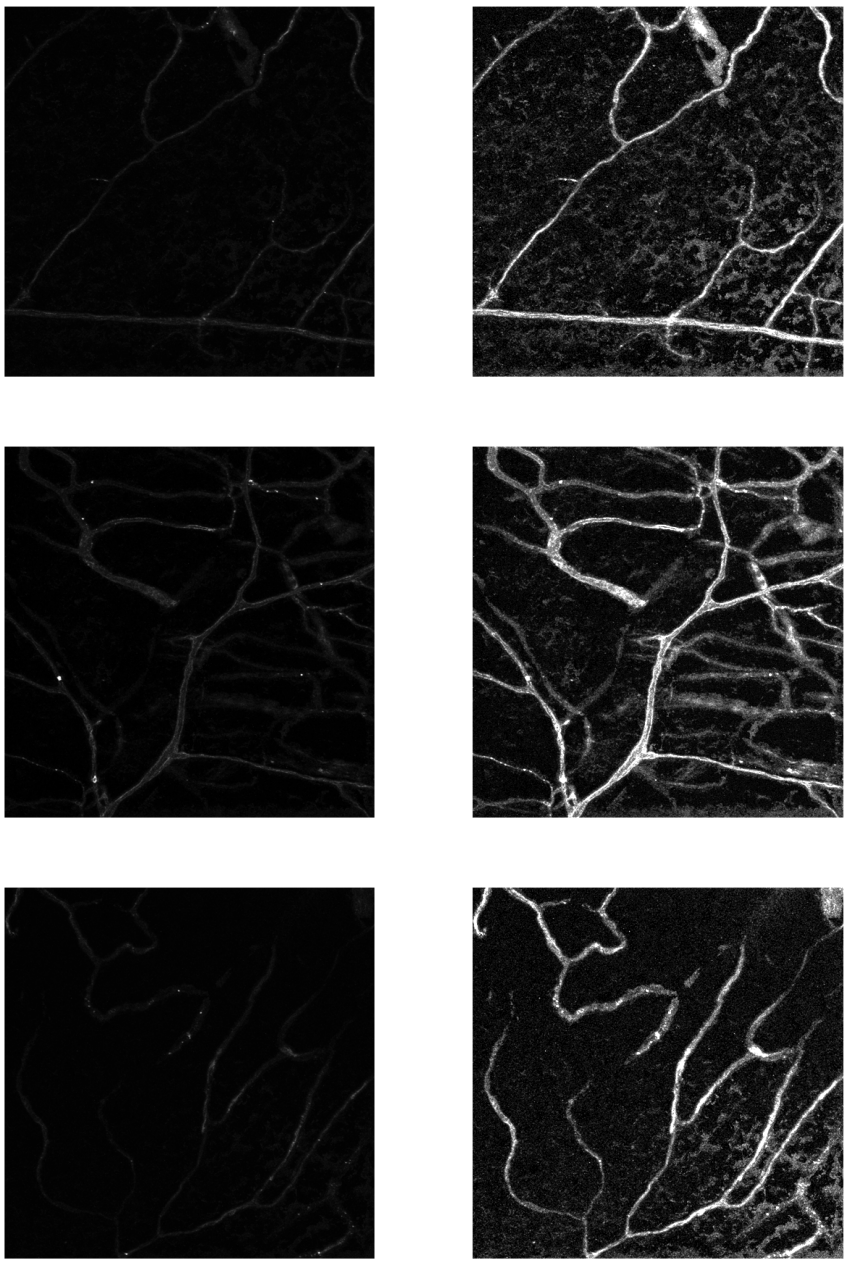


Figure 5.2. Sample confocal microscopy fused multi-focus images before (first column) and after (second column) adaptive histogram equalization.

ond subnetwork. The two final predictions from the DBAC and DDMAC networks are fused together to produce one final segmentation mask.

5.2.2.1 Deep Binary Attention Cascade (DBAC)

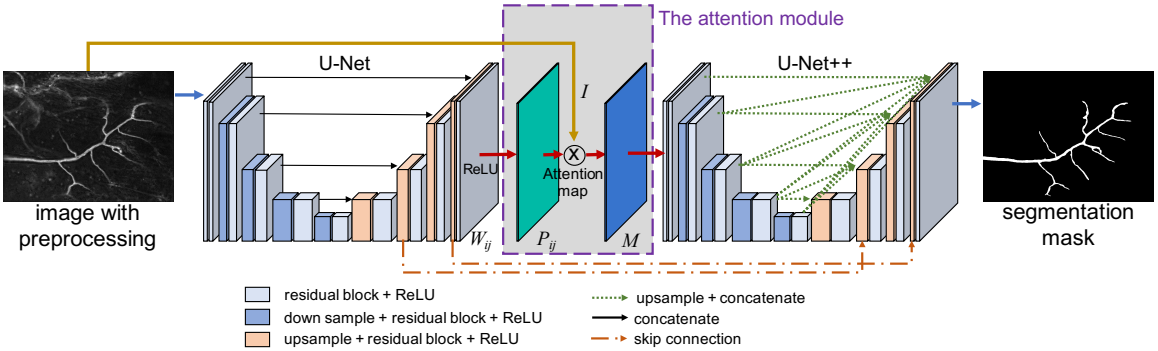


Figure 5.3. Architecture of the proposed deep binary attention cascade (DBAC).

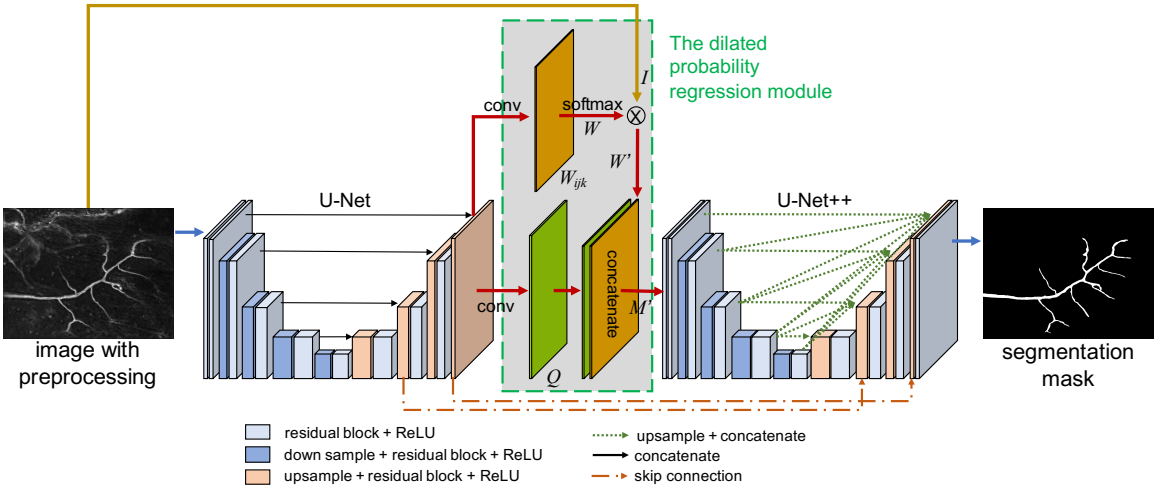


Figure 5.4. Architecture of the proposed deep distance map attention cascade (DDMAC).

The cascade network involves two subnetworks: a classical semantic segmentation network, UNet [26], followed by a deeper UNet++ [141] designed to exploit the convolutional features at different scales. The architecture of the proposed DBAC is shown in Figure 5.3. The convolutional layers have channels 32, 64, 128, 256, and 512 respectively from shallow

to deeper encoder layers in both U-Net and U-Net++. The decoder layers have symmetric number of channels. Each convolutional layer is replaced with a residual block [129].

The first subnetwork takes the input image and learns to predict a soft attention map where higher values indicate likelihood of vessel presence. The second subnetwork takes the input image and is guided by the soft attention map to performs coarse-to-fine refinement of the vessel segmentation mask from the first subnetwork.

The proposed attention module connecting the two subnetworks is defined as:

$$P_{ij} = max(0, W_{ij}) \tag{5.1}$$

$$M = I \times P \tag{5.2}$$

where W is the convolutional feature map from the first subnetwork, P is the 1-channel attention map prediction from the first subnetwork trained with a binary mask, where the positive values represent foreground and the negative values represent background. I is the input image, and M is the input to the second subnetwork.

We further extend this cascade by adding two shortcut connections that feed-forward the first subnetwork at different levels of the decoder to the corresponding levels of the decoder in the second subnetwork. The forwarded feature map and target feature map are concatenated together to prevent gradient vanishing as well as to directly forward feature maps from the first network to the second network.

The proposed network is trained on 60 epifluorescence microscopy images with resolution of 1360×1036 pixels [22][23], using binary cross-entropy loss [142] function and stochastic gradient descent (SGD) optimizer for the first and second subnetworks. Data augmentation is applied during the training and fine-tuning processes including random cropping (crop size is 448×448), rotation, flipping, scale, brightness, and contrast adjustment.

5.2.2.2 Deep Distance Map Attention Cascade (DDMAC)

As can be observed from Figure 5.1, cranial vasculature's diameter varies largely within even a small field of view. To better capture the varying size and shape of the vessels and to be able to detect thin vessels, we propose deep distance map attention cascade (DDMAC) shown in Figure 5.4. Dilated probability regression module (first subnetwork) aims to improve the precision score of the segmentation by applying multi-channels attention (the orange map in Figure 5.4) to the input image and then concatenating it with the dilated probability map Q (the green map in Figure 5.4).

Convolutional feature map Q is a 1-channel prediction from the first subnetwork of DDMAC and is trained with dilated probability map D, which is defined as:

$$D = distance\_transform(1 - V) \tag{5.3}$$

$$D = \{k|D>k\} \tag{5.4}$$

$$D = 1 - \frac{D}{max(D)} \tag{5.5}$$

where V represents the binary mask (vessel is positive, background is negative) used for training the proposed system. k is the upper bound of the pixel distance in D, which is set to be 20 in this study. The dilated probability regression module is defined as:

$$W = \frac{e^{W_{ijk}}}{\sum_{k=1}^{3} e^{W_{ijk}}} \tag{5.6}$$

$$W' = W \times I \tag{5.7}$$

$$M' = concatenate(W', Q) \tag{5.8}$$

where W_{ijk} is the convolutional feature map with size $i \times j \times k$ representing height, width, and channels respectively. In this study, i, j are the same as the height and width of the input image, and $k = 3$. The output M' is the input to the second subnetwork. The second

subnetwork takes M' and performs coarse-to-fine refinement of the vessel segmentation mask Q from the first subnetwork.

The proposed DDMAC network is trained on 60 epifluorescence microscopy images with resolution 1360×1036 [22][23] using the same data augmentation strategies as the described DBAC network. Mean square error (MSE) [142] and binary cross-entropy loss functions are used to train the first and the second subnetworks respectively. SGD is used as training optimizer.

5.2.3 Decision Fusion (Late Classifier Fusion)

For robust vessel segmentation performance, outputs of the two proposed networks DBAC and DDMAC are fused. Two decision fusion (late classifier fusion) mechanisms, average and maximum are considered:

$$Average: \qquad S_{ij} = \frac{L_{ij}+N_{ij}}{2} \tag{5.9}$$

$$Maximum: \qquad S_{ij} = max(L_{ij}, N_{ij}) \tag{5.10}$$

where L and N denote the 1-channel segmentation probability maps generated by the proposed DBAC and DDMAC networks respectively, and ij refers to pixel coordinates. Binary segmentation masks are produced by performing hysteresis thresholding [143] on the fused probability maps S. Mathematical morphology operations are applied to the predicted binary mask to fill small gaps and to remove small fragments.

5.3 Experimental Results

5.3.1 Data Collection

In this study, 7-8 weeks old C57BL/6J female mice were used to generate 80 3-D confocal microscopy image stacks. Following sacrifice, the entire body of the mouse was perfused through the heart with Kreb's/albumin solution containing AlexaFluor 594-conjugated soybean agglutinin (SBA) lectin to stain and identify blood microvessels [144].

Skull caps with dura mater were isolated and fixed in 10% formaldehyde. Images were acquired at 20x magnification on confocal FluoView FV1000 inverted microscope system (Olympus). The 160-270 mm thick Z-stacks were acquired with step size 1 mm. All animal experimental procedures were approved by the University of Missouri Institutional Animal Care and Use Committee.

The single focus Z-stacks microscopy images were fused using the method presented in [139] to produce eighty 512×512 multi-focus images. Out of those multi-focus images, 62 images were selected for training and 18 images were selected for testing. Silver truth (ST) segmentation masks were generated for all the training & test images by a combination of computer-generated segmentation masks and manual annotation. More precise ground truth (GT) segmentation masks were generated for the 18 test images by further inspection and manual correction by a domain expert.

5.3.2 Evaluation Metrics

Segmentation evaluations were carried out by sensitivity (recall), precision, specificity, accuracy, and dice score measures as defined below:

$$Sensitivity = \frac{TP}{TP + FN} \tag{5.11}$$

$$Precision = \frac{TP}{TP + FP} \tag{5.12}$$

$$Specificity = \frac{TN}{TN + FP} \tag{5.13}$$

$$Accuracy = \frac{TP + TN}{FN + FP + TP + TN} \tag{5.14}$$

$$Dice = \frac{2 * TP}{2 * TP + FP + FN} \tag{5.15}$$

where TP, TN, FP, FN represent number of true positive, true negative, false positive, and FN false negative pixels respectively.

Table 5.1. Number of trainable parameters of the compared networks.

Methods	Number of trainable parameters
U-Net++ [141]	71,106,794
deeper U-Net++	80,171,050
proposed DBAC	80,190,922
proposed DDMAC	80,210,861

5.3.3 Network Inference on 2-D Multi-focus Images

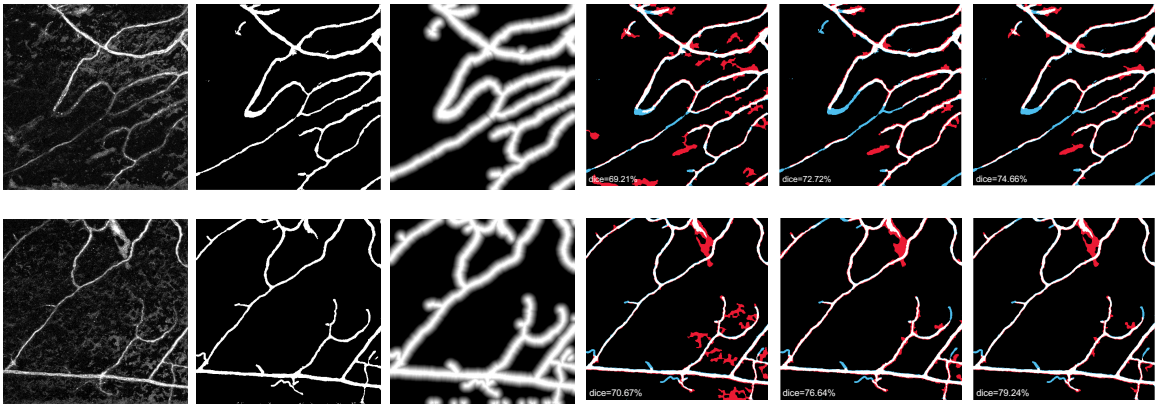
Because of limited amount of confocal microscopy training data, the proposed networks and their subnetworks were first trained with 60 epifluorescence microscopy images and their ground truth segmentation masks described in [22, 23]. Then the same networks were fine-tuned with 62 multi-focus confocal microscopy images and their associated silver truth masks described above. Both set of networks (with and without fine-tuning) were tested on the 18 multi-focus confocal microscopy test images and evaluated using corresponding ground truth segmentation masks.

5.3.3.1 Single Network Segmentation Performances

First, we compare segmentation performances of the proposed DBAC and DDMAC networks with state-of-the-art segmentation network UNet++ [141] and its deeper version with number of trainable parameters comparable to the proposed networks. The total number of trainable parameters of these networks are listed in Table 5.1.

The 1-channel probability maps outputted from the proposed DBAC and DDMAC networks are binarized using hysteresis thresholding [143] (lower bound=0.45, higher bound=0.95). Segmentation performances of the compared networks with and without fine-tuning are listed in Table 5.2. In order to better preserve the very thin vessels, up-sampled images (size 1024×1024) are inputted to the proposed DBAC network. Original sized images (512×512) are used with the proposed DDMAC network since regression to distance map is more robust to scale variations.

As we can see in Table 5.2, fine-tuning improves the dice score of all the deep learning networks by at least 1.4%. The proposed DBAC and DDMAC networks achieve the best sensitivity (recall) and precision scores respectively with and without fine-tuning. The distance based DDMAC network results in the best dice scores with and without fine-tuning.



(a) input image **(b)** attention module ground truth **(c)** dilated probability module ground truth **(d)** prediction from the DBAC with input size 1024×1024 **(e)** prediction from the DDMAC with input size 512×512 **(f)** final fused segmentation mask

Figure 5.5. Intermediate and final segmentation results for two sample multi-focus input images. (a) input images, (b-c) ground truth maps for the binary attention module in DBAC and dilated probability module in DDMAC, (d-f) predicted segmentation masks versus ground truth for DBAC, DDMAC, and ensemble networks. The red, blue, and white regions represent false-positive, false-negative, and true positive predictions respectively.

5.3.3.2 Ensemble Network Segmentation Performances

Table 5.3 summarizes segmentation performances of different configurations of the proposed ensemble network consisting of DBAC and DDMAC networks. The table explores two different input image sizes and two different decision fusion (late classifier fusion) mechanisms, average and maximum. All configurations of the proposed ensemble network outperform the best single network. Upsampling input images increases the precision scores. Ensemble network using average fusion and upscaled inputs for the proposed

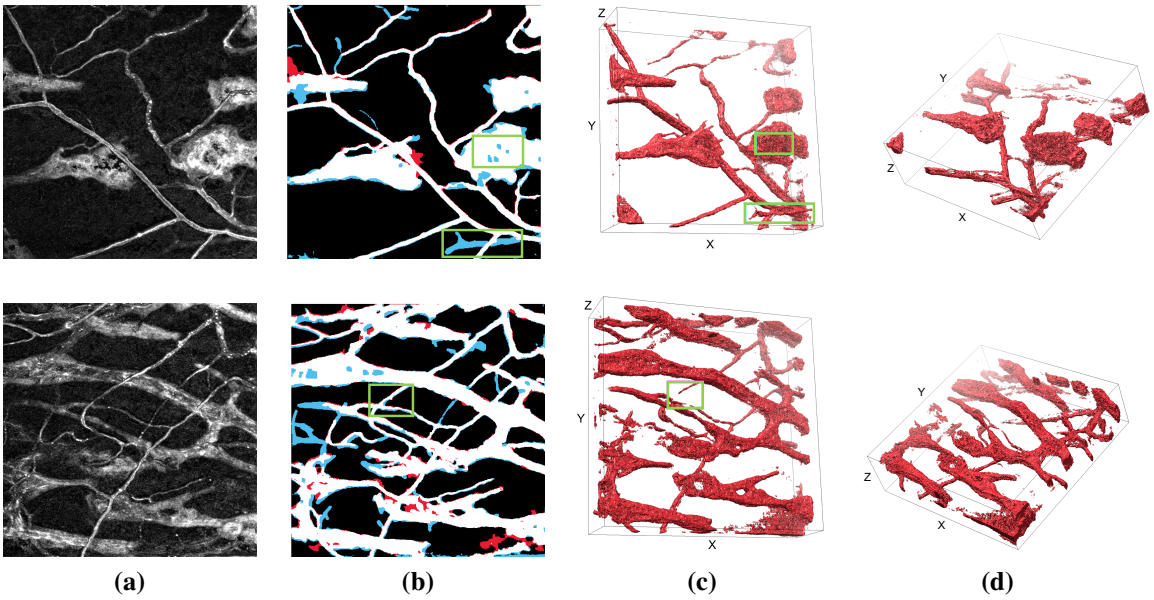


Figure 5.6. Sample outputs from the proposed system. (a) multi-focus confocal microscopy image enhanced with adaptive histogram equalization; (b) predicted segmentation mask versus ground truth where the red, blue, and white regions represent false-positive, false-negative, and true positive predictions respectively. (c-d) 3-D segmentation masks obtained by applying the proposed 2-D segmentation network to the individual single focus images forming the confocal microscopy volume. Visualization of the 3-D segmentation results were generated using the Chimera software [8]. Color fades with increasing depth.

DBAC network improves the dice score of the best performing single network by 2.29% reaching the best dice score of 81.45%.

Figure 5.5 shows single network and ensemble network segmentation results for two sample multi-focus images. Red and blue pixels represent false-positive and false-negative predictions respectively. False detections (false-positives) are typically caused by adverse effects of adaptive histogram equalization. Missed detections (false-negatives) are caused by thin vessels and low contrast between microvasculature and background. Fusing the outputs of the DBAC and DDMAC networks lead to recovery of some missed vessels and removal of some spurious detections (Figure 5.5f).

5.3.4 Network Inference on 3-D Image Stacks

The 3-D confocal microscopy image stacks in this study contain hundreds of slices. Each single slice captures the details of the specimen regions that lie close to its focal plane, while the remaining regions are imaged with poor contrast. Segmentation and visualization of the 3-D image stack allow comprehensive visualization and quantification of the anatomical structure of the 3-D microvasculature. 3-D deep learning networks could be employed to capture and learn the full 3-D morphological and anatomical characteristics of the microvasculature. However, demanding computational requirements and more importantly, the need for 3-D annotation for training limit their usability. In this section, we utilize the proposed 2-D image segmentation network to independently segment the single focus Z-stacks of a 3-D confocal microscopy volume.

Figure 5.6 shows 2-D multi-focus images, corresponding 2-D segmentation masks, and 3-D segmentation masks for two sample confocal microscopy volumes. The 2-D multi-focus images were obtained using the multi-focus fusion method described in [139]. 3-D segmentation masks were obtained by applying the proposed 2-D segmentation network to the individual single focus images forming the confocal microscopy volume. Each input slice has been preprocessed with linear contrast enhancement. Visualization of the 3-D segmentation results were generated using the Chimera software [8]. Figure 5.6c and 5.6d show promising 3-D segmentation results. In the first row of Figure 5.6, missed detections in 2-D (Figure 5.6b, blue pixels within green rectangles) caused by low contrast are recovered in 3-D (Figure 5.6c and 5.6d) thanks to linear contrast enhancement. In the second row of Figure 5.6, missed detections in 3-D (Figure 5.6c, green rectangle) caused by lack of semantic context in each slice can be corrected by the segmentation result of the corresponding 2-D multi-focus image (second row in Figure 5.6b).

5.4 Conclusion

In this chapter, we presented an ensemble of deep learning cascades for robust segmentation of blood vessels in confocal microscopy images. The proposed ensemble is composed of two complementary deep-learning cascades aiming to improve sensitivity and precision of the segmentation results. The proposed cascades first learn to predict two soft attention maps, one based on binary pixel classification, the other based on regression to a distance map. The attention maps guide the networks to predict an accurate vessel segmentation mask. Experiments demonstrated promising results towards segmention of microvasculatures in both 2-D and 3-D datasets. Segmentation is the key first step towards objective, and quantitative analysis of microvascular systems. The proposed segmentation system will be used to study microvascular remodeling.

Acknowledgement

This work is partially supported by awards from U.S.NIH National Institute of Neurological Disorders and Stroke R01NS110915. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the U. S. Government or agency thereof.

Table 5.2. Single network segmentation performances.

	Methods	Image size	Sensitivity %	Precision %	Specificity %	Accuracy %	Dice %
without fine-tuning	U-Net++ [141]	512×512	80.87	74.23	92.67	89.86	75.88
	deeper U-Net++	512×512	84.81	71.72	91.03	89.44	75.91
	proposed DBAC	512×512	**87.73**	66.78	88.39	88.00	74.22
	proposed DBAC	1024×1024	82.14	75.02	92.08	89.80	76.86
	proposed DDMAC	512×512	81.90	75.75	93.05	90.38	77.23
with fine-tuning	U-Net++ [141]	512×512	77.23	79.87	**95.00**	90.92	77.40
	deeper U-Net++	512×512	83.04	75.76	93.05	90.61	77.81
	proposed DBAC	512×512	84.41	76.02	92.72	90.79	78.66
	proposed DBAC	1024×1024	81.86	78.47	93.84	91.08	78.80
	proposed DDMAC	512×512	79.59	**80.12**	94.76	**91.50**	**79.16**

Table 5.3. Ensemble network segmentation performances.

DBAC with input size	DDMAC with input size	Fusion Mechanism	Sensitivity %	Precision %	Specificity %	Accuracy %	Dice %
512×512	512×512	maximum	**86.87**	73.61	91.47	90.44	79.69
512×512	512×512	average	81.95	79.87	94.41	91.67	80.90
1024×1024	512×512	maximum	86.30	74.63	92.08	90.75	80.04
1024×1024	512×512	average	80.60	**82.32**	**95.26**	**92.02**	**81.45**

CHAPTER 6

DISCRETE FOURIER TRANSFORM CLASS ACTIVATION MAP (DFT-CAM) AND WEAKLY-SUPERVISED SEGMENTATION

This chapter first introduces a Discrete Fourier Transform driven class activation map (DFT-CAM), then introduces the DFT-CAM based weakly-supervised object localization and segmentation systems.

We first introduce the DFT-CAM, a novel class activation map method that combines discrete Fourier transform based feature encoding with an orthogonality-based feature selection scheme. DFT-CAM doesn't require any training, better captures semantic information and aggregates only the most representative convolutional features. Besides, the proposed DFT-CAM can be applied to existing deep classification networks without changing the network architecture. To our best knowledge, the DFT-CAM is the first frequency-domain-based class activation map. Based on the DFT-CAM, we then proposed weakly-supervised object localization and segmentation systems using inexact data supervision strategies for 2-D image analysis.

In this chapter, we first introduce the background and basic idea of the class activation map and weakly-supervised learning. Then we introduce the details of the proposed DFT-CAM. Based on the DFT-CAM, we then proposed weakly-supervised object localization and segmentation systems. Experiments are conducted on both weakly-supervised object localization and segmentation systems using the ImageNet Large Scale Visual Recognition Challenge 2012 (ILSVRC2012) dataset [27] and human laryngeal endoscopy video data respectively. Finally, we conclude this chapter at the end.

6.1 Introduction

Deep learning has revolutionized computer vision tasks. Deep image classification networks such as CNN [11], VGG [9], residual network [129] and their variants; image segmentation networks such as fully convolutional network [145], U-Net [26], DeepLab [146] and their variants show promising results in many applications of image analysis and computer vision. While deep learning brings promising results in many fields, a significant problem is catching more and more attention, which is the need for large amounts of annotated data for deep learning training and testing.

Deep learning network training and testing require a huge amount of annotated data to learn from and evaluate, which is time-consuming and labor-intensive. In biomedical fields, the lack of annotated biomedical datasets is more severe than nature images, because of the difficulties in data acquisition and the expertise requirement for labeling. Weakly-supervised learning, a branch of machine learning, recently has received enormous attention, aiming to reduce data annotation workload and still keep the promising output precision. Based on the types of training data, weakly-supervised learning can be categorized into three types: (1) incomplete supervision, which means only a subset of training data is labeled; (2) inexact supervision, which means that the training data is coarse-grained; (3) inaccurate supervision, which means the training labels can be wrong [97]. We target inexact supervision in this chapter. Different methods have been proposed to attempt to achieve the goal, for example, methods that are based on specific system architectures and training strategies [98][96].

Recently, the class activation map (CAM) technique, which was originally developed for explainable AI, has been proposed to serve weakly-supervised learning. Basically, the CAM methods generate a discriminate saliency map for a specific class from the deep classification network, showing the pixel-wise probability of a pixel being used for the final class label prediction. A larger probability means a higher chance of the pixel belongs to the target object. This transformation from image-level (class label) to pixel-level (class

activation map) enables pixel-wise precision output by only using image-level ground truth labels, which can reduce a lot of data annotation workload. Therefore, CAM techniques have recently been applied to weakly-supervised learning tasks such as weakly-supervised object localization (WSOL) [98][96] and weakly-supervised object segmentation [147], as they bring up a possible way of reducing the annotation workload of training data.

Several CAM methods have been proposed, such as gradient-free methods CAM [91] and Ablation-CAM [92], or gradient-based methods such as Grad-CAM [93] and Grad-CAM++ [94]. To generate a saliency map, these methods often combine information from all the channels from a convolutional layer using a weighted sum operation. This process can blend unrelated regions of the target object and affect the energy distribution of the saliency map, thus, lowering the accuracy of the downstream tasks.

The vocal folds (VFs) are a pair of muscles in the larynx, the tubular structure that connects the throat to the windpipe (trachea). The VFs function like a valve in the upper airway, opening, and closing as needed for breathing, swallowing, and speaking [108]. Therefore, VF dysfunction can cause breathing difficulty (dyspnea), swallowing dysfunction (dysphagia), and/or voice impairment (dysphonia), all of which can significantly reduce the patient's quality of life, even cause life-threatening situations [110] [111].

Flexible Endoscopic Evaluation of Swallowing with Sensory Testing (FEESST) is a clinical test used by speech-language pathologists (SLPs) and otolaryngologists to examine motor and sensory functions of the VFs and to assess the risk of aspiration [114]. The FEESST procedure involves passing a thin flexible endoscope through the nose into the pharynx and larynx to visualize the VFs. Small puffs of air are delivered through the endoscope to stimulate the laryngeal mucosa near the VFs, triggering the laryngeal adductor reflex (LAR). During this airway protective reflex, the VFs abruptly close momentarily (less than 1 second) to prevent invasion of "foreign" material into the lungs. However, the FEESST-generated videos (if they are even recorded) are only visually inspected, resulting in the loss of potentially clinically valuable information to facilitate diagnosis and

guide treatment planning. Besides, vocal fold motion analysis is needed to detect subtle VF dysfunction that may be missed by visual inspection alone, as well as to objectively monitor disease progression or treatment response. The current practice of vocal fold motion analysis mainly relies on fully-supervised deep-learning semantic segmentation, which requires labor-intensive and time-consuming pixel-level annotation of the glottal region in endoscopy images [148] [149] [6]. Also, most of the proposed glottal region segmentation methods are designed for high-speed transoral videoendoscopy, which has a high frame rate, resolution, and image quality. In this chapter, we focus on the more challenging analysis of VF videos obtained by transnasal flexible endoscopy, which has a low frame rate, resolution, and image quality.

In this chapter, we first proposed a Discrete Fourier Transform driven class activation map method named DFT-CAM. Then, based on the DFT-CAM, we proposed a weakly-supervised object localization system.

6.2 Related Works

6.2.1 Weakly-supervised Learning in Biomedical

Recently, weakly-supervised deep learning has caught tons of attention due to its ability to reduce the human workload of data annotation and improve training efficiency. Weakly-supervised learning can be categorized into three main categories: incomplete, inaccurate, and inexact learning [97]. Weakly-supervised learning has been adapted for the biomedical application of deep learning, aiming to relieve the lack of annotated biomedical data, such as [150] for pharyngeal phase analysis and [151] for tumor lesions segmentation. To our best knowledge, we are the first team to develop weakly-supervised learning to segment the glottal region in a low-speed transnasal laryngeal endoscopy video.

6.2.2 Laryngeal Endoscopy Video Analysis

Studying the vocal fold motion in a laryngeal endoscopy video is one of the crucial ways to extract objective and quantitative information that can be used to analyze vocal fold dysfunction, which can cause dyspnea, dysphagia, and dysphonia. Considering the output data modalities, analysis related to vocal fold motion can be divided into segmentation-based methods and classification-based methods. Segmentation-based methods are usually intended to segment the vocal fold related regions from the input images, for example, the glottis and two pairs of vocal fold muscles. The segmented areas are then used to further extract motion data such as supraglottic index and glottic anterior angle [47] [5] [152] [153]. Among them, deep learning based methods require manual-annotated pixel-level labels for training, which is time-consuming. These segmentation-based methods are mainly designed for high-speed videoendoscopy, where VFs are fully visible. They often detect spurious regions when VFs are fully-closed, occluded, or are not in the field of view of the endoscope, which usually happened in the application of low-speed transnasal endoscopy videos. Current classification-based methods [30] [25] [154] are designed to classify each low-speed transnasal laryngeal endoscopy image into one of three classes: non-LAR (open VFs), LAR (closed VFs), and visual occlusion (where the VFs are either obstructed by other anatomical structures or are out of the endoscope camera field of view). This type of method only needs image-level labels of the vocal fold, which can reduce human workload. However, image-level labels can only provide limited details about the vocal fold motion.

While a great improvement over the previous works, both segmentation-based and classification-based methods have flaws in their applications. In order to tackle these problems, we first proposed a weakly-supervised glottal region segmentation network to reduce the annotation workload and provide detailed vocal fold motion information. Then, we further proposed a multi-task laryngeal analysis system of low-speed transnasal endoscopy video at the end of this chapter, which concatenates the glottal region segmentation results

with image-based vocal fold classification results to improve the overall robustness of the laryngeal endoscopy video analysis.

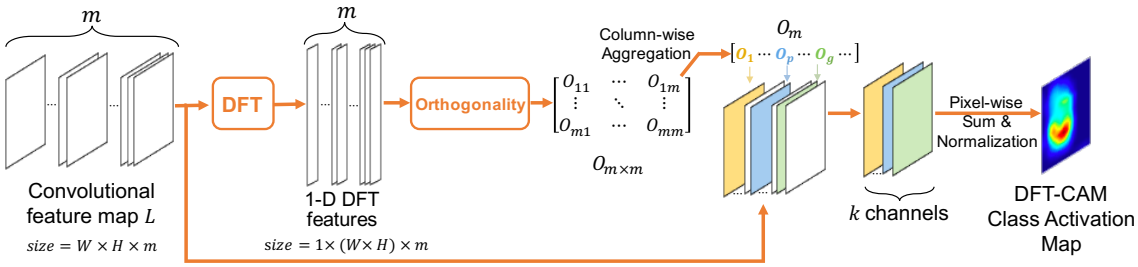


Figure 6.1. Processing steps of the proposed discrete Fourier transform driven class activation map DFT-CAM. The figure illustrates selection of k representative channels out of m original channels through DFT-based feature encoding and orthogonality-based feature selection. Selected feature channels are then aggregated to produce the proposed class activation map.

6.3 Method of DFT-CAM

In this section, we propose a gradient-free, discrete Fourier transform driven, class activation map method named DFT-CAM. DFT-CAM is inspired by the hierarchical feature learning capabilities of deep classification networks and the effectiveness of frequency-domain representations in image compression [155]. Discrete Fourier Transform (DFT) converts spatial domain information in the images into the frequency domain. Frequency domain representation allows better separation of significant semantic information from image details and noise. These representations can be used to summarize geometrical characteristics of spatial information [24]. The proposed DFT-CAM method first uses discrete Fourier transform (DFT) based representation to summarize learned features in convolutional feature maps; then uses feature orthogonality to automatically select the most representative semantic features while preventing the inclusion of less-contributed features. Feature orthogonality has been shown to be effective in region selection [25] and hyper-parameters selection [156][157] in deep learning. The proposed two-step DFT and orthogonality-based approach result in a more accurate class activation map of target foreground objects.

Without loss of generality, we use a classical 2D convolutional neural networks (CNN) [11] to demonstrate the proposed DFT-CAM algorithm. We pick the last convolutional layer Q located before the first fully-connected layer of the CNN. Given an m-channel convolutional feature map L outputted from the convolutional layer Q, the proposed DFT-CAM is computed as follows.

Step-1: Discrete Fourier transform (DFT) [24] is computed for each $W \times H$ feature map at channel L_i:

$$F_i[s,t] = \frac{1}{WH} \sum_{w=1}^{W} \sum_{h=1}^{H} L_i[w,h] e^{-j2\pi\left(\frac{s}{W}w + \frac{t}{H}h\right)} \tag{6.1}$$

The output is converted to a $1 \times n$ DFT feature vector where $n = W \times H$. Concatenation of channel level feature vectors results in an $m \times n$ feature matrix $\mathcal{F}_{mn}$. $\mathcal{F}_{mn}$ serves as a convolutional feature summary of the $W \times H \times m$ convolutional feature map L.

Step-2: Using $\mathcal{F}_{mn}$ and dot product operation, we can compute the orthogonality between each pair of convolutional channels of L using the corresponding 1-D flattened DFT feature vectors as follows:

$$O_{mm} = \mathcal{F}_{mn} \cdot \mathcal{F}_{mn} - diag(\mathcal{F}_{mn} \cdot \mathcal{F}_{mn}) \tag{6.2}$$

Here, we eliminate the diagonal elements from the dot product results to remove the orthogonality between a single channel and itself. O_{mm} is the output orthogonal matrix with size $m \times m$, where each row represents the orthogonalities between a single channel and other channels. The orthogonality (also called uniqueness) of each feature channel L_i is computed by the sum of the rows in O_{mm} as follows:

$$O_i = \sum_{j=1}^{m} O_{mm}(i,j) \tag{6.3}$$

Step-3: We first determine indices of the feature maps with top k orthogonality values:

$$\mathcal{K} = \underset{i \in \{1,..,n\}}{\text{Top}\, k}(O_i) \tag{6.4}$$

Then, the k feature maps corresponding to the index set $\mathcal{K}$ are aggregated to generate a single channel response as follows:

$$R(x, y) = \sum_{i \in \mathcal{K}} L_i(x, y) \tag{6.5}$$

Following the common activation operation in Grad-CAM [93], Ablation-CAM [92], and Grad-CAM++ [94], we apply a rectified linear unit (ReLU) [158] and normalization after the previous steps to generate the final DFT-CAM output, so as to only keep the positive part of the CAM and normalize the class activation map to $[0, 1]$. This process is defined as:

$$D = \text{ReLU}(R) \tag{6.6}$$

$$\text{DFT-CAM} = \frac{D - min(D)}{max(D) - min(D)} \tag{6.7}$$

The proposed DFT-CAM adopts DFT encoding to summarize learned convolutional features from a specific convolutional layer. The summarized information is then used to select the most representative convolutional channels through a maximum orthogonality criterion to form the final CAM. The proposed DFT-CAM can be easily applied to existing deep learning networks for generating CAM visualization without modification of the architecture.

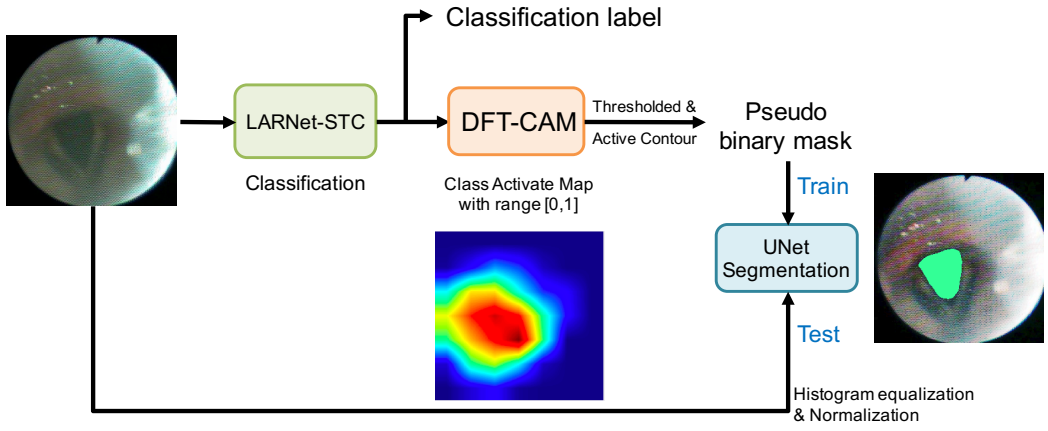


Figure 6.2. Weakly-supervised glottal region segmentation.

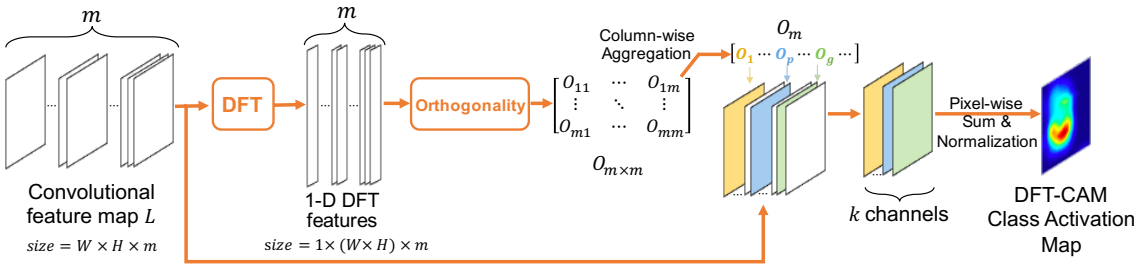


Figure 6.3. Processing steps of the proposed Discrete Fourier Transform driven class activation map DFT-CAM. The figure illustrates selection of k representative channels out of m original channels through DFT-based feature encoding and orthogonality-based feature selection. Selected feature channels are then aggregated to produce the proposed class activation map.

6.4 Method of DFT-CAM Based Weakly-supervised Glottal Region Segmentation

In this section, we describe the technical details of the proposed weakly-supervised glottal region segmentation. The whole system is shown in Figure 6.2. The proposed pipeline involves three main modules:

1. Deep classification module. This module uses our previously proposed deep classification network LARNet-STC [25] to predict an image class label for the input image.

2. Pseudo binary mask generation. Our proposed Discrete Fourier Transform driven class activation map (DFT-CAM) is used to generate a class activation map from the LARNet-STC during the inference. The generated CAM is then processed by classical image-processing techniques to produce a pseudo binary mask of the glottal region.

3. Weakly-supervised segmentation. A two-outputs UNet [26] is trained using all the pseudo binary masks to segment the glottal region from the input image.

Details about each module are introduced in the following sub-sections.

6.4.1 Deep Classification Module

Previously, we have developed two deep classification networks for human laryngeal closure detection. One is using spatial information of the transnasal endoscopy image, called LARNet [30]. Another one is built on top of the LARNet, which is using spatial-temporal information of the video, named LARNet-STC [25]. These two networks are trained and tested on human transnasal endoscopy images using image-level labels, aiming at classifying each image into one of three classes: non-LAR (open VFs), LAR (closed VFs), and visual occlusion (the VFs are either masked/covered by other anatomical structures or out of the camera field of view). In this chapter, we will use LARNet-STC to generate convolutional maps for the input transnasal endoscopy image.

6.4.1.1 Classical Image-processing Based Pseudo Binary Mask Generation

By applying the DFT-CAM, we have generated a class activation map (CAM) for the input transnasal endoscopy image. The generated CAM is a map of pixel-wise probabilities, where each map pixel has a probability about itself being used for the class label prediction. The larger the probability, the more likely it is that the pixel belongs to the target object. Since the CAM is a pixel-level prediction, we can use the CAM to generate a pseudo segmentation mask for the glottal region in the input image. First, we threshold the DFT-CAM output and generate a coarse segmentation mask. From this coarse segmentation mask, we apply Active contour [24] to generate a pseudo binary mask of glottal region, where positive values indicate glottal region and zeros indicate background region. Morphological methods such as binary erosion and dilation are applied for further refinement. The quality of the pseudo segmentation masks are depending on the CAM and the Active contour accuracy.

6.4.2 Weakly-supervised Glottal Region Segmentation

Since the pseudo segmentation masks are generated using the error-prone CAM and the Active contour, the accuracy of the pseudo segmentation masks is changing between

different input images, depending on image quality and deep classification network performance. We want the segmentation network to learn from good pseudo segmentation masks and correct those poor segmentation results.

After we generated pseudo segmentation masks for all images, we train an UNet [26] to predict a binary mask for the glottal region in an input image. The UNet architecture is shown in Figure 6.4. This UNet has two prediction outputs, one is a binary mask for the glottal region, and the other one is a binary mask of the edge of the glottal region. To train this UNet, we generate two kinds of datasets, one is the pseudo binary masks generated directly from Section 6.4.1.1, and the other one is glottal region edge masks, which are generated by subtracting the binary erosion of the pseudo binary mask from the pseudo binary mask.

Besides, the inaccurate pseudo segmentation masks will confuse the deep segmentation network during training and thus lowers the accuracy of the deep segmentation network. To tackle this problem, we design a training data filtering strategy. During the first 20 epochs, the UNet learns from all the available training data. Starting from the $21st$ epochs, for every 10 epochs, we check for each training image to see if its glottal region mask prediction accuracy is lower than a certain threshold. If it satisfied the condition, we will remove this image in the following training epochs. In the experiment, we set the threshold to be 60% of the average prediction Intersection over Union (IoU) of the previous training epoch.

Both prediction outputs are trained with a combination of Binary Cross Entropy loss and Dice loss. Optimizer is Adam optimizer. The number of training epochs is a hundred.

6.5 Experimental Results of Weakly-supervised Object Localization

In this section, we evaluate the proposed DFT-CAM method in terms of its weakly-supervised object localization performance. Weakly-supervised object localization (WSOL) aims to localize objects in an image using only image-level class labels. The first CAM-

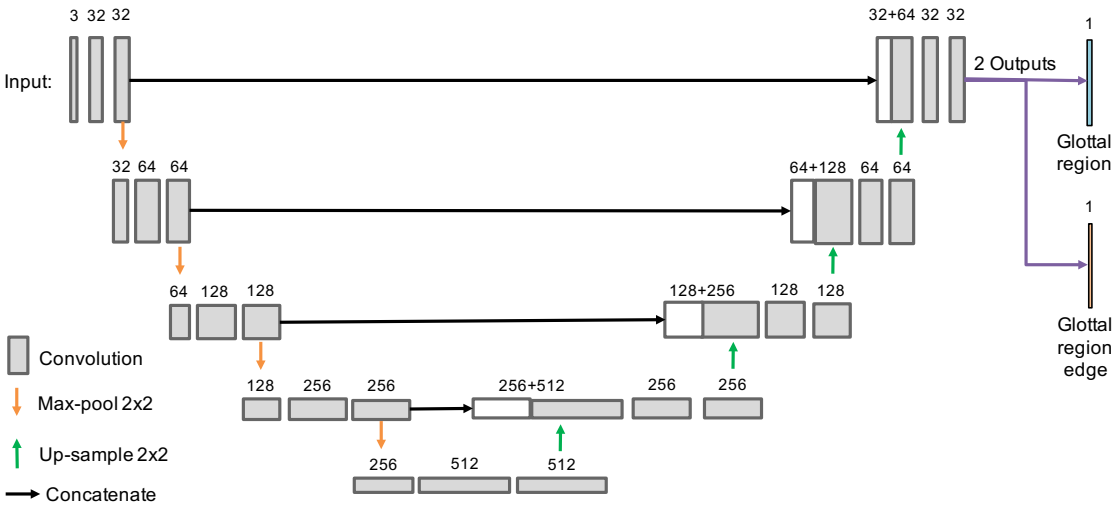


Figure 6.4. UNet architecture with two prediction outputs, one is a binary mask for the glottal region, and the other one is a binary mask for the edge of the glottal region. Numbers above the convolutional blocks are the corresponding numbers of convolutional channels.

based WSOL approach has been proposed in [91]. The process consists of the computation of class activation maps from deep classification networks, followed by thresholding of the obtained class activation maps to generate binary segmentation masks or object bounding boxes. These WSOL schemes attract great attention because they reduce labor-intensive annotation needs.

6.5.1 Dataset

All evaluations and comparisons in this study have been performed using the validation set of the ImageNet Large Scale Visual Recognition Challenge 2012 (ILSVRC2012) dataset [27]. The ILSVRC2012 dataset contains 1.2 million training images of 1,000 classes. The validation set contains 50,000 images with associated ground truth class labels and object bounding boxes.

Table 6.1. Average Intersection over Union (IoU) (%) scores between the ground truth and the CAM-generated bounding boxes. Bold fonts mark the best results, underlined fonts mark the second-best results.

Methods	AlexNet [159]	VGG-16 [9]	ResNet-101 [129]	Inception_v3 [160]
Grad-CAM [93]	<u>50.30</u>	50.85	**50.29**	**48.72**
Grad-CAM++ [94]	47.96	49.05	48.39	48.09
Ablation-CAM [92]	48.44	50.69	48.49	48.29
Eigen-gradcam [161]	39.24	40.78	47.60	47.68
Layer-CAM [162]	47.45	48.89	48.18	48.07
Conv-CAM (ours)	48.09	<u>51.25</u>	<u>49.42</u>	<u>48.64</u>
DFT-CAM (ours)	**51.06**	**54.22**	48.67	**48.72**

6.5.2 Weakly-supervised Object Localization (WSOL)

We evaluated the proposed DFT-CAM method in terms of its weakly-supervised object localization (WSOL) performance and compared to the recent state-of-the-art CAM methods including Grad-CAM [93], Grad-CAM++ [94], Ablation-CAM [92], Eigen-gradCAM [161], and Layer-CAM [162]. Two main innovations of the proposed DFT-CAM approach are (1) feature encoding using discrete Fourier transform, and (2) orthogonality-based feature selection. In order to further assess the role of DFT in DFT-CAM performance, we have built another CAM method named Conv-CAM. Conv-CAM relies on the same orthogonality-based feature selection scheme but performs feature selection on a raw convolutional feature map without DFT encoding. The experiments were conducted on four deep learning classification networks, AlexNet [159], VGG-16 [9], ResNet-101 [129], and Inception_v3 [160]. The object bounding boxes were generated as follows:

1. Classification networks (AlexNet [159], VGG-16 [9], ResNet-101 [129], and Inception_v3 [160]) were directly loaded from PyTorch [163] models library with their corresponding pre-trained weights.
2. CAM methods were applied to the feature maps outputted from the last convolutional layer (preceding the fully-connected layers) of the selected deep networks.

3. All CAMs outputs $C(i, j)$ were binarized to generate binary segmentation masks M_C:

$$M_C(i, j) = \begin{cases} 1 \text{ if } C(i, j) > 0.15 \times max(C) \\ 0 \text{ otherwise} \end{cases}$$

(this process was applied only to network outputs with correct class predictions)

4. Object bounding boxes $\mathcal{B}$ were computed around the largest connected components of the binary masks M_C.

In this experiment, we only considered the top-1 classification results for all the networks, and top $k = 5$ feature channels for the DFT-CAM and Conv-CAM methods.

6.5.3 Evaluation of Weakly-supervised Object Localization

We evaluated the CAM results in terms of intersection over union scores, $IoU(\mathcal{B}_C, \mathcal{B}_G) = |\mathcal{B}_C \cap \mathcal{B}_G| \,/\, |\mathcal{B}_C \cup \mathcal{B}_G|$, between the CAM-generated ($\mathcal{B}_C$) and the ground truth ($\mathcal{B}_G$) bounding boxes. The IoU scores for the proposed DFT-CAM, Conv-CAM, and the other state-of-the-art CAM methods are listed in Table 6.1. As we can see in Table 6.1, the proposed DFT-CAM results in the best IoU scores on three of the deep learning networks (four networks in total) including AlexNet [159], VGG-16 [9], and Inception_v3 [160].

When the two proposed feature orthogonality-based CAM methods are compared, the DFT-CAM with discrete Fourier transform (DFT)-based feature encoding outperforms the Conv-CAM, demonstrating the importance of DFT encoding. Meanwhile, the Conv-CAM achieves the 2nd best results in VGG-16 [9], ResNet-101 [129], Inception_v3 [160], and the 3rd best result in AlexNet [159], illustrating the effectiveness of orthogonality in convolutional feature selection.

Figure 6.5 shows sample WSOL results for different CAM methods using VGG-16 [9] classification network on the ILSVRC2012 validation set [27]. As we can see in Figure 6.5, the proposed DFT-CAM method generates more accurate CAMs for target objects. For example, in Figure 6.5 on the 2nd row, Grad-CAM++ [94], Ablation-CAM [92], and Layer-CAM [162] highlight both foreground and background regions, whereas, the pro-

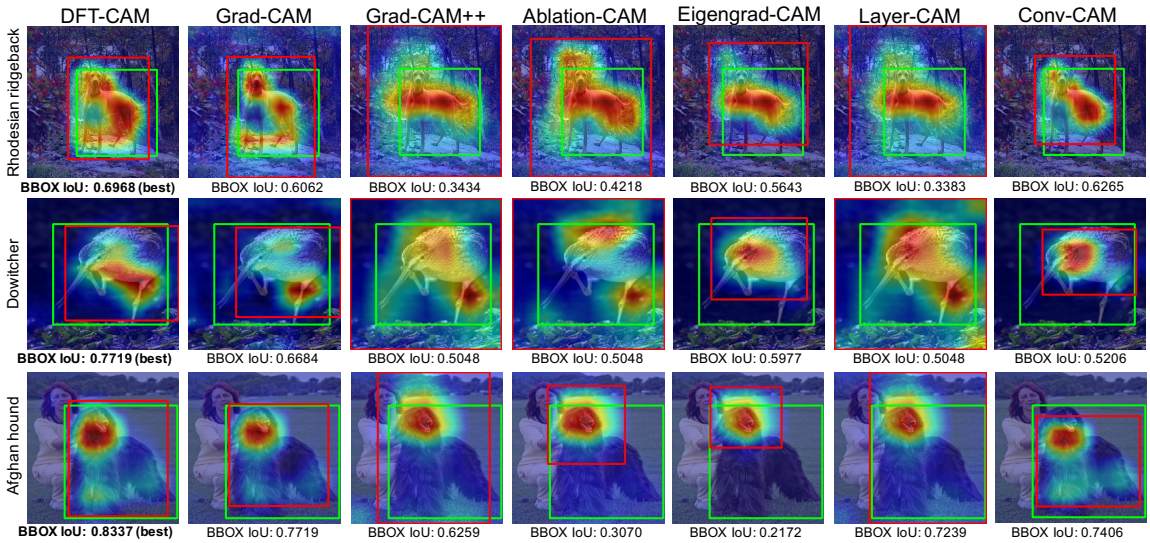


Figure 6.5. Sample weakly-supervised object localization results obtained from the VGG-16 classification network [9] outputs using the different class activation map (CAM) methods. The heatmaps overlayed on the original images illustrate the CAM results. The red to blue colormap represents high to low probabilities. All bounding boxes were generated from the class activation maps as described in Section 6.5.2. Each column shows the outputs from the same CAM method. Each row shows the same input image and its class label. The green and red bounding boxes correspond to the ground truth and CAM bounding boxes respectively. The number below each image represents the IoU score between the ground truth and CAM boxes.

Table 6.2. DFT-CAM weakly-supervised object localization IoU scores (%) for different k values.

	AlexNet [159]	VGG-16 [9]	ResNet-101 [129]	Inception_v3 [160]
$k = 1$	47.03	49.86	**48.89**	48.57
$k = 3$	50.38	53.17	48.70	**48.72**
$k = 5$	**51.06**	**54.22**	48.67	**48.72**

posed DFT-CAM method better highlights only the target object parts. Also, in Figure 6.5 on the 3rd row, Ablation-CAM [92] and Eigen-gradCAM [161] mainly highlight the head of the Afghan hound but miss part of the body, as the comparison, the proposed DFT-CAM detects the whole body of the target object. The experimental results demonstrate the better capabilities of the proposed DFT-CAM method for the localization of target objects under different circumstances.

6.5.4 Ablation Study of DFT-CAM

We analyzed the performance of the proposed DFT-CAM method for different number of selected feature channels k.

As we can see in Table 6.2, for AlexNet [159], VGG-16 [9], and Inception_v3 [160] networks, the IoU score increases when k increases. For the ResNet-101 [129] network, the IoU score decreases as k increases. This can be due to the residual connections in the ResNet-101 which can learn sparse and unique features in different channels of the last convolutional layer. This results in smaller orthogonality between channels of the last convolutional layer and is harder to find the most representative target object features. Meanwhile, background features can be brought into the CAM output as the k value increases, thus lowering the IoU score of the WSOL task.

6.6 Experimental Results of Weakly-supervised Glottal Region Segmentation

Table 6.3. Average Intersection over Union (IoU) (%) scores between the ground truth and the predicted bounding boxes of the glottal region. Bold fonts mark the best results.

CAM methods	CAM threshold values	Active contour (Average %)	UNet segmentation (Average %)
Grad-CAM	0.4	25.76	37.98
Grad-CAM	0.5	31.50	36.65
Grad-CAM	0.6	34.55	32.24
Grad-CAM	0.7	32.02	42.76
Grad-CAM	0.8	22.11	47.17
Grad-CAM	0.9	14.01	21.94
DFT-CAM (ours)	0.4	36.13	43.68
DFT-CAM (ours)	0.5	**37.32**	**68.32**
DFT-CAM (ours)	0.6	32.83	65.54

6.6.1 Dataset

Our video dataset was collected with a 3.7-mm outer-diameter endoscope with a 1.5-mm inner-diameter working channel (11302BD2, Karl Storz), at a frame rate of 30 FPS

(frames per second). The endoscope tip was positioned at a typical level for viewing laryngeal pathology to permit visualization of the bilateral VFs throughout the procedure. Twenty healthy nonsmoking human subjects (7 men and 13 women) aged 20 to 40 years were recruited and tested. In total 58 videos were collected, and each video frame is with size $480 \times 720 \times 3$. The data collection protocol was approved by the University of Missouri Institutional Review Board. We randomly separate 58 videos into training and testing videos. The training set has 46 videos (79.3%). The testing set has 12 videos (20.7%). The training set is only used to train the LARNet-STC [25] for classification.

Original video frames extracted from the raw endoscopy videos come with a black region surrounding the round visual area. We generate a tight bounding box around the visual area and crop the black region outside the bounding box. The cropped images are then resized to 224×224 following the setting in LARNet-STC paper [25].

Two ground truth datasets are used in this experiment for evaluation. One is the glottal region bounding box, which is generated for all the frames in the testing video set. The other one is the glottal region segmentation mask, which is manually annotated only for two testing videos due to the timing issue. The glottal region segmentation mask has positive values for the foreground and zeros for the background.

6.6.2 Weakly-supervised Glottal Region Segmentation

The experiment is conducted as follows:

1. The trained LARNet-STC [25] is used to predict a classification label for each frame in the testing video set.

2. Class activation map method is applied to a specific convolutional layer in the LARNet-STC [25] during the inference. In this experiment, the selected convolutional layer is the last convolutional layer of ResNet-18 from the original-image stream in LARNet, which is located at the third input image feature extraction stream in LARNet-STC.

3. Active contour and morphological methods are applied to the thresholded class activation map to generate pseodo binary mask M for the glottal region. Additionally, masks of glottal region edge E will be generated by subtracting the binary erosion of the pseudo binary mask from the pseudo binary mask ($E = M - erosion(M)$).

4. All generated glottal region and glottal region edge pseodo binary masks are used to train the two-outputs UNet. All masks have positive values for the foreground and zeros for the background. The input images of the two-outputs UNet are histogram equalized video frames, the same as LARNet-STC [25].

5. The trained UNet will be used to predict glottal region segmentation mask for the testing video set. If multiple segments are predicted for the glottal region in the same image, we will only keep the one with the maximum average probability. A bounding box will be generated surrounding the positive region in the final mask.

As a comparison to our proposed DFT-CAM, we use Grad-CAM [93] to generate class activation maps in Step 2 mentioned above. Other steps stay the same. For simplification, we called the two-outputs UNet trained with DFT-CAM outputs "UNet-DFT", and the two-outputs UNet trained with Grad-CAM outputs "UNet-Grad".

6.6.3 Quantitative Evaluation of Segmentation

We evaluated the segmentation results in terms of intersection over union scores (IoU) for both bounding box and segmentation mask ground truth, $IoU(\mathcal{B}_P, \mathcal{B}_G) = |\mathcal{B}_P \cap \mathcal{B}_G| \,/\, |\mathcal{B}_P \cup \mathcal{B}_G|$, between the prediction ($\mathcal{B}_P$) and the ground truth ($\mathcal{B}_G$).

First, we evaluate our proposed weakly-supervised glottal region segmentation pipeline using glottal region bounding box ground truth. The average IoU scores are listed in Table 6.3. As we can see in Table 6.3, our proposed DFT-CAM with threshold 0.5 reaches the best average IoU regarding the bounding boxes evaluation, for both Active contour and UNet, showing the effectiveness of our proposed DFT-CAM in locating the region of the target

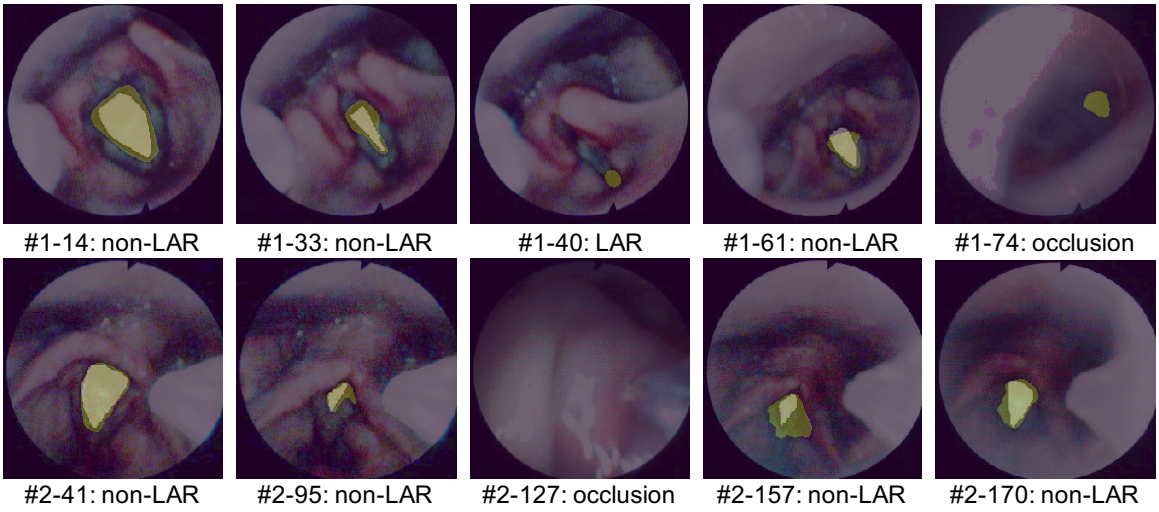
Table 6.4. Average Intersection over Union (IoU) (%) scores between the ground truth and the predicted glottal region masks. Bold fonts mark the best results.

CAM threshold values	Methods	Active contour (Average %)	UNet (Average %)
0.8	Grad-CAM	27.41	47.48
0.5	DFT-CAM (ours)	**48.54**	**54.74**

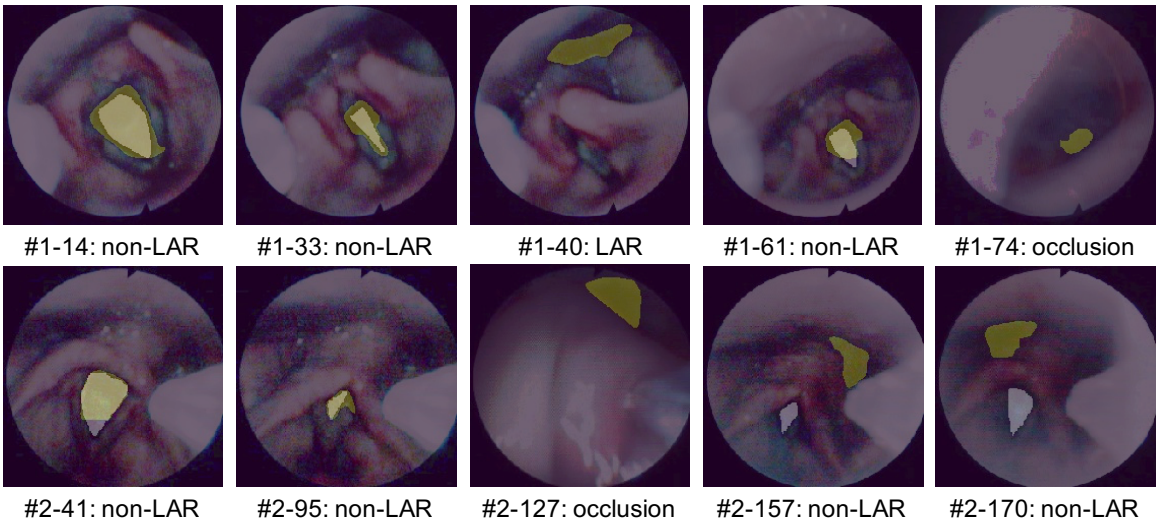
object. As a comparison to our proposed DFT-CAM, we set different threshold values ranging from 0.4 to 0.9 for the Grad-CAM and generated their corresponding bounding boxes from segmentation masks generated by Active contour and UNet. However, the best average IoU score of Grad-CAM at the Active contour column is 34.55%, which is 2.77% lower than the best score of DFT-CAM. Also, the best average IoU score of Grad-CAM at the UNet segmentation column is 47.17%, which is 21.15% lower than the best score of DFT-CAM.

Secondly, by comparing the IoU scores of Active contour and UNet segmentation, we can see that by training the two-outputs UNet using pseudo binary masks generated from the Active contour, the average IoU scores of the UNet received remarkable improvements for both Grad-CAM and the proposed DFT-CAM compared to Active contour, except for the one that generated by Grad-CAM with a threshold value 0.6. Regarding our proposed DFT-CAM with a threshold value 0.5, applying the weakly-supervised segmentation method increases the bounding box accuracy from 37.32% to 68.32%, which is an $\sim$ 83% improvement compared to the IoU score of Active contour of DFT-CAM.

Finally, we evaluate our proposed weakly-supervised segmentation pipeline on manually-annotated glottal region masks ground truth. This ground truth set is generated only for two videos, and these two videos contain frames belonging to non-LAR (open VFs), LAR (closed VFs), and visual occlusion (the VFs are either masked/covered by other anatomical structures or out of the camera field of view) classes. UNet-DFT is following the same



(a) Predictions from the UNet-DFT



(b) Predictions from the UNet-Grad

Figure 6.6. Visualization comparison of UNet-DFT and UNet-Grad predictions. The yellow transparent mask is the UNet prediction. The solid grey mask is the glottal region segmentation ground truth mask. The index below each image indicates "#video sequence index - frame index: class label" of that image. (a) contains segmentation outputs predicted by the two-outputs UNet that is trained using DFT-CAM's outputs. (b) contains segmentation outputs predicted by the two-outputs UNet that is trained using Grad-CAM's outputs. Some of the frames don't have glottal region segmentation ground truth masks because they are either LAR (VFs closed) or occluded, and the vocal fold is not visible in these cases.

setting as before. From Table 6.3 we can see that, the best IoU score of the UNet-Grad is produced by using a threshold value of 0.8, and the best IoU score of the UNet-DFT is produced by using a threshold value of 0.5. Thus, in this evaluation, we generate glottal region segmentation masks using two-outputs UNet that trained with Grad-CAM outputs with a threshold value of 0.8 and DFT-CAM outputs with a threshold value of 0.5, respectively. The IoU scores comparison is shown in Table 6.4. As we can see in Table 6.4, our proposed DFT-CAM reaches the best IoU in both Active contour and UNet methods, presenting the robustness of the DFT-CAM in locating the target region using convolutional features. Also, IoU scores of both UNet-DFT and UNet-Grad are increased compared to the Active contour column, presenting the powerfulness of the proposed weakly-supervised segmentation pipeline.

In the end, we compared our proposed UNet-DFT (CAM threshold=0.5) with other deep-learning glottal region segmentation methods. We directly used code from U-LSTM [5], OpenHSV [153], and Hamad et al. [6]. U-LSTM [5], OpenHSV [153] are designed and trained on high-speed trans-oral laryngeal endoscopy video. Hamad et al. [6] is developed for low-speed trans-nasal laryngeal endoscopy video analysis, which has the same type of images as ours. The input images are from the two manually-annotated videos. The ground truth dataset we use for this comparison includes glottal region bounding boxes and manually-annotated segmentation binary masks. The comparison is shown in Table 6.5.

As we can see in Table 6.5, deep learning networks that are designed for high-speed trans-oral laryngeal videoendoscopy are not feasible for our low-speed trans-nasal laryngeal endoscopy, which result in low IoU scores. Our proposed UNet-DFT achieves the best compared to other methods, showing that although our proposed method is not fully-supervised on glottal region segmentation mask ground truth, it can still predict promising results.

Table 6.5. Average Intersection over Union (IoU) (%) scores comparison between multiple glottal region segmentation methods

Method	Glottal region bounding box (Avg. %)	Glottal region binary mask (Avg. %)
U-LSTM [5]	9.45	3.23
OpenHSV [153]	5.71	2.55
Hamad et al. [6]	46.17	29.50
UNet-DFT (ours, CAM threshold=0.5)	**57.94**	**54.74**

6.6.4 Qualitative Evaluation of Segmentation

Qualitative comparisons of predictions from UNet-DFT and UNet-Grad are shown in Figure 6.6. The yellow transparent mask is the UNet output. The grey solid mask is the glottal region segmentation ground truth mask. The index below each image indicates "#video sequence index - frame index: class label" of that image. As we can see in Figure 6.6, predictions from UNet-DFT are more precisely compared to the ones with UNet-Grad. Especially for occlusion and LAR cases, where the vocal fold is not visible, UNet-DFT predicts small or no false positive regions. However, UNet-Grad predicts large false positive regions. Note that the two images #2-157 and #2-170 in Figure 6.6 are out-of-focus and with low resolution, the UNet-DFT can still detect regions at the glottal area, however, the UNet-Grad detects other regions outside the glottal area. These comparison results show the robustness of the proposed weakly-supervised segmentation pipeline in the glottal region segmentation task.

6.6.5 Multi-task Analysis for Human Laryngeal Trans-nasal Endoscopy Video Data

Furthermore, we proposed a multi-task laryngeal analysis system of low-speed transnasal endoscopy video which concatenates both image-level and pixel-level information. For each input image, we generate both image-level and pixel-level labels. The image-level labels are predicted by the deep classification network [25] used at the beginning of the

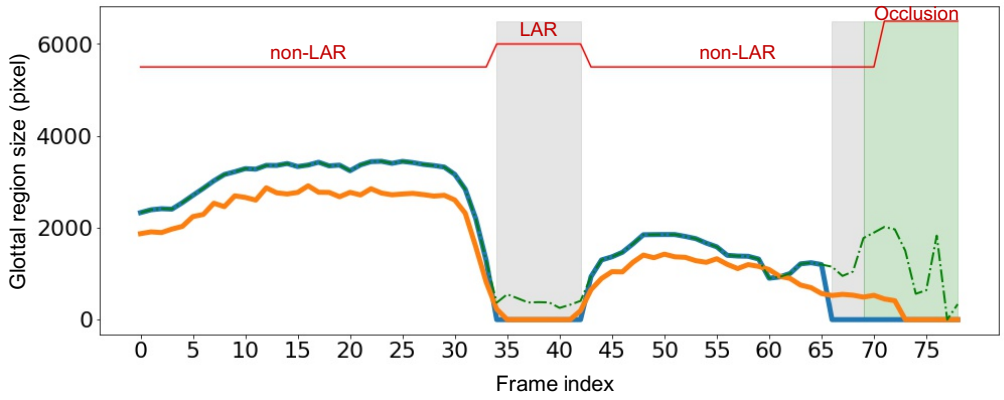


Figure 6.7. Output of the multi-task laryngeal analysis system for a low-speed transnasal endoscopy video. The X-axis is the video frame index. The Y-axis is the glottal region size. The blue signal represents the classification result-corrected glottal region sizes. The orange signal represents ground truth glottal region sizes. The green dash represents the UNet segmented glottal region sizes. The red signal represents the ground truth classification label, where the low signal is "non-LAR" class, the median-high signal is "LAR" class, and the highest signal is "occlusion" class (note that the ground truth classification label doesn't reflect the glottal region size of the corresponding frame). The grey region represents the predicted "LAR" class. The green region represents the predicted "occlusion" class. The white region besides grey and green represents the predicted "non-LAR" class.

weakly-supervised glottal region segmentation. The pixel-level labels are generated by the weakly-supervised glottal region segmentation network. The multi-task laryngeal analysis system combines the image-level and pixel-level predictions of the input video sequence, enabling a comprehensive analysis of the laryngeal endoscopy video. The multi-task laryngeal analysis system has the following advantages:

1. The proposed system, which predicts both image-level and pixel-level labels, only requires image-level ground truth labels for training that can save a great amount of annotation workload compared to pixel-level labels.

2. The system predicts both image-level and pixel-level labels for an input image. The image-level label can help distinguish the all-zero mask prediction between LAR

class (closed VFs) and visual occlusion (where the VFs are either obstructed by other anatomical structures or are out of the endoscope camera field of view).

3. The system provides richer details about the vocal fold motion compared to classification-based laryngeal endoscopy analysis [30] [25] [154].

A sample output of the multi-task laryngeal analysis system is shown in Figure 6.7. In Figure 6.7, the X and Y axes represent the video frame index and glottal region size. The orange signal represents ground truth glottal region sizes. The green dash represents the UNet segmented glottal region sizes. The blue signal represents the glottal region sizes corrected by the corresponding LARNet-STC [25] classification labels of LAR and occlusion classes. The red signal represents the ground truth classification label, where the low signal is the "non-LAR" class, the median-high signal is the "LAR" class, and the highest signal is the "occlusion" class. Note that the ground truth classification label doesn't reflect the glottal region size of the corresponding frame. The grey and green regions represent the predicted "LAR" and "occlusion" classes. The white region besides grey and green represents the predicted "non-LAR" class.

By simultaneously analyzing results generated by the classification-based and segmentation-based methods of laryngeal endoscopy video, the overall robustness of the laryngeal endoscopy video analysis is improved and more comprehensive information about the input laryngeal endoscopy video is provided.

6.7 Conclusion

Deep learning has been applied to many fields and produced promising results. However, training deep learning networks usually requires a huge amount of annotated data. Data annotation is time-consuming and labor-intensive. Specifically, in the biomedical field, annotated data is even harder to acquire compared to nature images because of privacy concerns and expertise requirements.

In this chapter, we proposed a Discrete Fourier Transform Driven Class Activation Map (DFT-CAM) and DFT-CAM based weakly-supervised object localization and segmentation systems, which utilize image-level labels for training and are able to predict pixel-level labels. The DFT-CAM is a novel discrete Fourier transform (DFT) driven class discrimination map. The proposed method uses DFT to better encode the semantic information captured in each feature channel and uses feature orthogonality criterion to select the most representative convolutional features. This scheme improves the overall WSOL accuracy by preventing the inclusion of less-essential convolutional features to the class activation map. Promising results were obtained for WSOL task across different deep learning classification networks.

Then a weakly-supervised object segmentation system is proposed. The proposed system first uses the LARNet-STC network to predict a classification label for the input image. Then, DFT-CAM is applied to generate a class activation map from the LARNet-STC during the inference. The generated class activation maps are processed with classical image-processing methods to produce pseudo binary masks of glottal regions. Finally, a two-outputs UNet trained with all pseudo binary masks will predict a glottal region segmentation mask for the input image. To further improve the robustness and comprehensiveness of the laryngeal endoscopy video analysis, we proposed a multi-task analysis system, which concatenates the classification and segmentation results together.

Experimental results conducted on both object localization and segmentation tasks show the robustness of the proposed pipelines.

Acknowledgement

This work is partially supported by awards from US NIH NINDS R01NS110915. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the U. S. Government or agency thereof.

CHAPTER 7

CONCLUSION

In this dissertation, we proposed novel deep-learning solutions involving various attention mechanisms, supervision manners, information selection, and fusion methods, deep-learning visualization approaches, and network architectures. The proposed algorithms enable the analysis of different data modalities involving 1-D signal + time, 2-D image and 3-D image volume, and 2-D image + time (video), as well as their associated biomedical signal, images, and video analysis problems.

For 1-D signal + time data analysis, we proposed DeepDDK and multi-modal multi-scale DeepDDK that allow weakly supervised learning for oral-diadochokinesis syllable events detection in audio files using syllable timestamps training data, which reduces lots of manual annotation of oral-DDK syllables compared to other deep learning based oral-DDK solutions. For 2-D image and 3-D image volume analysis, we proposed the ensemble of deep learning cascades using global soft attention mechanisms aimed at the improvement of recall and precision respectively. These methods were applied to the segmentation of meningeal vascular networks in confocal microscopy images. The proposed 2-D image segmentation system allows weakly supervised learning using a very small amount of training data. For 2-D + time (video) analysis, we proposed LARNet and LARNet-STC using global hard attention mechanisms, which allow rare and short-term video events detection in laryngoscopy videos without segmenting and tracking the target object in the video and thereby considerably reducing manual annotation workload. Within the LARNet and LARNet-STC, the proposed orthogonal region selection network (ORS) can extract the re-

gion of interest from an image using global hard attention in an unsupervised manner and improve the overall classification accuracy. Furthermore, explainable artificial intelligence visualization has been applied to justify the effectiveness of our proposed LARNet and LARNet-STC networks, as well as to improve the overall reliability and interpretability of our proposed deep learning solutions. Besides, for explainable-AI and weakly-supervised learning, we proposed Discrete Fourier Transform driven class activation map (DFT-CAM), which is a gradient-free method to generate a class activation map from deep classification networks. The proposed DFT-CAM was applied to LARNet and LARNet-STC to visualize the spatial regions in the input images that led to correct classification response. The proposed DFT-CAM method can be applied to arbitrary deep classification networks without changing the network architecture. Based on the DFT-CAM method, we further proposed weakly-supervised object localization and segmentation systems. These detection and segmentation systems do not require expensive manual annotation of object bounding boxes or segmentation masks for training.

In this dissertation, we dealt with three different types of data modalities associated with four different biomedical data analysis tasks. In 1-D signal + time data analysis, we focused on speech and swallow studies involving oral-diadochokinesis (oral-DDK) audio data. In 2-D image analysis, we focused on meningeal vascular system studies involving confocal microscopy. In 2-D + time (video) analysis, we focused on vocal folds motion studies involving vocal folds laryngoscopy video data.

By developing novel deep learning based solutions, we enabled automated, objective, and quantitative analysis of multi-modal multi-dimensional biomedical data. Experimental results showed robustness and promising results for the proposed deep learning algorithms regarding different biomedical datasets. Furthermore, the proposed explainable artificial intelligence techniques increased the reliability and improved interpretability of our proposed deep learning networks, and enabled weakly-supervised learning for reducing data annotation workload. Accurate, objective, and quantitative analysis of biomedical data is of

great significance because these analyses can be potentially used in early diagnosis, disease progress monitoring, and treatment development.

This dissertation can be further extended by incorporating the proposed methods and tools into a cross-platform, multi-modal, and multi-dimensional biomedical data analysis system, which combines information from all involved data modalities to provide a comprehensive analysis of neurological disorders.

BIBLIOGRAPHY

1. O. Räsänen, G. Doyle, and M. Frank, "Pre-linguistic segmentation of speech into syllable-like units," *Cognition*, vol. 171, pp. 130–150, 2018.

2. Z. Smekal, J. Mekyska, I. Rektorova, and M. Faundez-Zanuy, "Analysis of neurological disorders based on digital processing of speech and handwritten text," in *IEEE Int. Symposium on Signals, Circuits and Systems*, 2013, pp. 1–6.

3. J. Mekyska, Z. Smekal, M. Kostalova, M. Mrackova, S. Skutilova, and I. Rektorova, "Motor aspects of speech imparment in Parkinson's disease and their assessment," *Ceska A Slovenska Neurologie A Neurochirurgie*, vol. 74, no. 6, pp. 662–668, 2011.

4. Y. Y. Wang, K. Gao, A. M. Kloepper, Y. Zhao, M. Kuruvilla-Dugdale, T. E. Lever, and F. Bunyak, "Deepddk: A deep learning based oral-diadochokinesis analysis software," in *IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, 2019, pp. 1–4.

5. M. K. Fehling, F. Grosch, M. E. Schuster, B. Schick, and J. Lohscheller, "Fully automatic segmentation of glottis and vocal folds in endoscopic laryngeal high-speed videos using a deep convolutional lstm network," *PLoS ONE*, vol. 15, no. 2, p. e0227791, 2020.

6. A. Hamad, M. Haney, T. E. Lever, and F. Bunyak, "Automated segmentation of the vocal folds in laryngeal endoscopy videos using deep convolutional regression networks," in *Proc. IEEE Conf. Comp. Vision and Pattern Recog. Workshops*, 2019.

7. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," *Int. J. Computer Vision*, vol. 128, no. 2, pp. 336–359, 2020.

8. E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin, "UCSF Chimera—a visualization system for exploratory research and analysis," *J. Computational Chemistry*, vol. 25, no. 13, pp. 1605–1612, 2004.

9. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

10. I. R. I. Haque and J. Neubert, "Deep learning approaches to biomedical image segmentation," *Informatics in Medicine Unlocked*, vol. 18, p. 100297, 2020.

11. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.

12. S. R. Stahlschmidt, B. Ulfenborg, and J. Synnergren, "Multimodal deep learning for biomedical data fusion: a review," *Briefings in Bioinformatics*, vol. 23, no. 2, p. bbab569, 2022.

13. S. Dong, P. Wang, and K. Abbas, "A survey on deep learning and its applications," *Computer Science Review*, vol. 40, p. 100379, 2021.

14. G. Haskins, U. Kruger, and P. Yan, "Deep learning in medical image registration: a survey," *Machine Vision and Applications*, vol. 31, pp. 1–18, 2020.

15. S. Budd, E. C. Robinson, and B. Kainz, "A survey on active learning and human-in-the-loop deep learning for medical image analysis," *Medical Image Analysis*, vol. 71, p. 102062, 2021.

16. Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," *Neurocomputing*, vol. 452, pp. 48–62, 2021.

17. A. de Santana Correia and E. L. Colombini, "Attention, please! a survey of neural attention models in deep learning," *Artificial Intelligence Review*, vol. 55, no. 8, pp. 6037–6124, 2022.

18. S. Chaudhari, V. Mithal, G. Polatkan, and R. Ramanath, "An attentive survey of attention models," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 12, no. 5, pp. 1–32, 2021.

19. P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable ai: A review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, p. 18, 2020.

20. S. Nazir, D. M. Dickson, and M. U. Akram, "Survey of explainable artificial intelligence techniques for biomedical imaging with deep neural networks," *Computers in Biology and Medicine*, p. 106668, 2023.

21. E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (xai): Toward medical xai," *IEEE Trans. Neural Networks and Learning Systems*, vol. 32, no. 11, pp. 4793–4813, 2020.

22. Y. M. Kassim, O. V. Glinskii, V. V. Glinsky, V. H. Huxley, G. Guidoboni, and K. Palaniappan, "Deep u-net regression and hand-crafted feature fusion for accurate blood vessel segmentation," in *IEEE Intl. Conf. Image Processing (ICIP)*, 2019, pp. 1445–1449.

23. Y. M. Kassim, V. S. Prasath, R. Pelapur, O. V. Glinskii, R. J. Maude, V. V. Glinsky, V. H. Huxley, and K. Palaniappan, "Random forests for dura mater microvasculature segmentation using epifluorescence images," in *Annual Intl. Conf. IEEE Engineering in Medicine and Biology Society (EMBC)*, 2016, pp. 2901–2904.

24. J. C. Russ, *The Image Proc. Handbook*. CRC press, 2016.

25. Y. Y. Wang, A. S. Hamad, K. Palaniappan, T. E. Lever, and F. Bunyak, "Larnet-stc: Spatio-temporal orthogonal region selection network for laryngeal closure detection in endoscopy videos," *Computers in Biology and Medicine*, vol. 144, p. 105339, 2022.

26. O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Int. Conf. on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.

27. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and F. Li, "ImageNet Large Scale Visual Recognition Challenge," *Intl. J. Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

28. Y. Y. Wang, K. Gaol, A. Hamad, B. McCarthy, A. M. Kloepper, T. E. Lever, and F. Bunyak, "Multi-modal and multi-scale oral diadochokinesis analysis using deep learning," in *IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, 2021, pp. 1–6.

29. Y. Y. Wang, O. Glinskii, F. Bunyak, and K. Palaniappan, "Ensemble of deep learning cascades for segmentation of blood vessels in confocal microscopy images," in *IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, 2021, pp. 1–7.

30. Y. Y. Wang, A. S. Hamad, T. E. Lever, and F. Bunyak, "Orthogonal region selection network for laryngeal closure detection in laryngoscopy videos," in *42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2020, pp. 2167–2172.

31. J. Godino-Llorente, S. Shattuck-Hufnagel, J. Choi, L. Moro-Velázquez, and J. Gómez-García, "Towards the identification of Idiopathic Parkinson's Disease from the speech. new articulatory kinetic biomarkers," *PLoS ONE*, vol. 12, no. 12, 2017.

32. H. Zhang, A. Wang, D. Li, and W. Xu, "DeepVoice: A voiceprint-based mobile health framework for Parkinson's disease identification," in *IEEE Int. Conf. Biomedical & Health Informatics (BHI)*, 2018, pp. 214–217.

33. C. Poellabauer, N. Yadav, L. Daudet, S. Schneider, C. Busso, and P. Flynn, "Challenges in concussion detection using vocal acoustic biomarkers," *IEEE Access*, vol. 3, pp. 1143–1160, 2015.

34. J. Rusz, B. Benova, H. Ruzickova, M. Novotny, T. Tykalova, J. Hlavnicka, T. Uher, M. Vaneckova, M. Andelova, K. Novotna *et al.*, "Characteristics of motor speech phenotypes in multiple sclerosis," *Multiple Sclerosis and Related Disorders*, vol. 19, pp. 62–69, 2018.

35. M. Horne, L. Power, and D. Szmulewicz, "Quantitative assessment of syllabic timing deficits in ataxic dysarthria," in *IEEE Engineering in Medicine and Biology Society (EMBC)*, 2018, pp. 425–428.

36. Y. Wang, R. Kent, J. Duffy, and J. Thomas, "Analysis of diadochokinesis in ataxic dysarthria using the motor speech profile program™," *Folia Phoniatrica et Logopaedica*, vol. 61, no. 1, pp. 1–11, 2009.

37. F. Tao, L. Daudet, C. Poellabauer, S. Schneider, and C. Busso, "A portable automatic PA-TA-KA syllable detection system to derive biomarkers for neurological disorders." in *Interspeech*, 2016, pp. 362–366.

38. A. Kumar and B. Raj, "Deep CNN framework for audio event recognition using weakly labeled web data," *arXiv preprint arXiv:1707.02530*, 2017.

39. G. Son, S. Kwon, and Y. Lim, "Speech rate control for improving elderly speech recognition of smart devices," *Advances in Electrical and Computer Engineering*, vol. 17, no. 2, pp. 79–85, 2017.

40. J. Pons, R. Gong, and X. Serra, "Score-informed syllable segmentation for a cappella singing voice with convolutional neural networks," *arXiv preprint arXiv:1707.03544*, 2017.

41. T. Koç and T. Çiloğlu, "Automatic segmentation of high speed video images of vocal folds," *J. Applied Mathematics*, vol. 2014, 2014.

42. Y. Zhang, E. Bieging, H. Tsui, and J. J. Jiang, "Efficient and effective extraction of vocal fold vibratory patterns from high-speed digital imaging," *J. Voice*, vol. 24, no. 1, pp. 21–29, 2010.

43. T. Shi, H. J. Kim, T. Murry, P. Woo, and Y. Yan, "Tracing vocal fold vibrations using level set segmentation method," *Int. J. Numerical Methods in Biomed. Engr.*, vol. 31, no. 6, p. e02715, 2015.

44. H. I. Turkmen, M. E. Karsligil, and I. Kocak, "Classification of laryngeal disorders based on shape and vascular defects of vocal folds," *Computers in Biology and Medicine*, vol. 62, pp. 76–85, 2015.

45. C.-F. J. Kuo, Y.-H. Chu, P.-C. Wang, C.-Y. Lai, W.-L. Chu, Y.-S. Leu, and H.-W. Wang, "Using image processing technology and mathematical algorithm in the automatic selection of vocal cord opening and closing images from the larynx endoscopy video," *Comp. Methods and Prog. in Biomedicine*, vol. 112, no. 3, pp. 455–465, 2013.

46. M. M. Haney, A. Hamad, E. Leary, F. Bunyak, and T. E. Lever, "Automated quantification of vocal fold motion in a recurrent laryngeal nerve injury mouse model," *The Laryngoscope*, vol. 129, no. 7, 2019.

47. J. Lin, E. S. Walsted, V. Backer, J. H. Hull, and D. S. Elson, "Quantification and analysis of laryngeal closure from endoscopic videos," *IEEE Trans. Biomed. Engr.*, vol. 66, no. 4, pp. 1127–1136, 2019.

48. S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

49. A. Bochkovskiy, C. Wang, and H. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

50. C. Deng, M. Wang, L. Liu, Y. Liu, and Y. Jiang, "Extended feature pyramid network for small object detection," *IEEE Trans. Multimedia*, p. 1, 2021.

51. M. Henaff, A. Szlam, and Y. LeCun, "Recurrent orthogonal networks and long-memory tasks," *arXiv preprint arXiv:1602.06662*, 2016.

52. E. Vorontsov, C. Trabelsi, S. Kadoury, and C. Pal, "On orthogonality and learning recurrent networks with long term dependencies," in *Proc. Int. Conf. on Machine Learning*, 2017, pp. 3570–3578.

53. L. Huang, X. Liu, B. Lang, A. W. Yu, Y. Wang, and B. Li, "Orthogonal weight normalization: Solution to optimization over multiple dependent stiefel manifolds in deep neural networks," in *AAAI Conf. on Artificial Intelligence*, 2018.

54. A. Prakash, J. Storer, D. Florencio, and C. Zhang, "Repr: Improved training of convolutional filters," in *Proc. IEEE Conf. Comp. Vision and Pattern Recog.*, 2019, pp. 10 666–10 675.

55. W. Hu, L. Xiao, and J. Pennington, "Provable benefit of orthogonal initialization in optimizing deep linear networks," in *Int. Conf. Learning Representations*, 2019.

56. L. Xiao, Y. Bahri, J. Sohl-Dickstein, S. S. Schoenholz, and J. Pennington, "Dynamical isometry and a mean field theory of cnns: How to train 10,000-layer vanilla convolutional neural networks," *arXiv preprint arXiv:1806.05393*, 2018.

57. J. Lezama, Q. Qiu, P. Musé, and G. Sapiro, "Ole: Orthogonal low-rank embedding-a plug and play geometric loss for deep learning," in *Proc. IEEE Conf. on Comp. Vision and Pattern Recog.*, 2018.

58. Y. Wang, D. Gong, Z. Zhou, X. Ji, H. Wang, Z. Li, W. Liu, and T. Zhang, "Orthogonal deep features decomposition for age-invariant face recognition," in *Proc. European Conf. Comp. Vision*, 2018, pp. 738–753.

59. D. Masko and P. Hensman, "The impact of imbalanced training data for convolutional neural networks," 2015.

60. H. Lee, M. Park, and J. Kim, "Plankton classification on imbalanced large scale database via convolutional neural networks with transfer learning," in *IEEE Int. Conf. on Image Processing*, 2016.

61. S. Pouyanfar, Y. Tao, A. Mohan, H. Tian, A. S. Kaseb, K. Gauen, R. Dailey, S. Aghajanzadeh, Y.-H. Lu, S.-C. Chen *et al.*, "Dynamic sampling in convolutional neural networks for imbalanced data classification," in *IEEE Conf. Multimedia Inf. Proc. Retrieval*, 2018, pp. 112–117.

62. C. Lin, S. Chen, P. S. Santoso, H. Lin, and S. Lai, "Real-time single-stage vehicle detector optimized by multi-stage image-based online hard example mining," *IEEE Trans. Vehicular Technology*, vol. 69, no. 2, pp. 1505–1518, 2019.

63. S. Wang, W. Liu, J. Wu, L. Cao, Q. Meng, and P. J. Kennedy, "Training deep neural networks on imbalanced data sets," in *Int. joint Conf. on Neural Networks*, 2016, pp. 4368–4374.

64. T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE Trans. Pat. Analysis Machine Intel.*, vol. 42, no. 02, pp. 318–327, 2020.

65. Q. Dong, S. Gong, and X. Zhu, "Imbalanced deep learning by minority class incremental rectification," *IEEE Trans. Pat. Analysis Machine Intel.*, vol. 41, no. 6, pp. 1367–1381, 2018.

66. T. Wu, Q. Huang, Z. Liu, Y. Wang, and D. Lin, "Distribution-balanced loss for multi-label classification in long-tailed datasets," in *Proc. European Conf. Comp. Vision*. Springer, 2020, pp. 162–178.

67. S. H. Khan, M. Hayat, M. Bennamoun, F. A. Sohel, and R. Togneri, "Cost-sensitive learning of deep feature representations from imbalanced data," *IEEE Tran. Neural Networks and Learning Systems*, vol. 29, no. 8, pp. 3573–3587, 2017.

68. H. Wang, Z. Cui, Y. Chen, M. Avidan, A. B. Abdallah, and A. Kronzer, "Predicting hospital readmission via cost-sensitive deep learning," *IEEE/ACM Trans. Computational Biology and Bioinformatics*, vol. 15, no. 6, pp. 1968–1978, 2018.

69. M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, vol. 106, pp. 249–259, 2018.

70. F. A. Gers, N. N. Schraudolph, and J. Schmidhuber, "Learning precise timing with lstm recurrent networks," *J. Machine Learning Research*, vol. 3, no. Aug, pp. 115–143, 2002.

71. R. Dey and F. M. Salem, "Gate-variants of gated recurrent unit (gru) neural networks," in *60th IEEE Intl. Midwest Symposium on Circuits and Systems (MWSCAS)*, 2017, pp. 1597–1600.

72. Z. Mao, Y. Su, G. Xu, X. Wang, Y. Huang, W. Yue, L. Sun, and N. Xiong, "Spatio-temporal deep learning method for adhd fmri classification," *Information Sciences*, vol. 499, pp. 1–11, 2019.

73. M. Bengs, N. Gessert, M. Schlüter, and A. Schlaefer, "Spatio-temporal deep learning methods for motion estimation using 4d oct image data," *Int. J. Computer Assisted Radiology and Surgery*, vol. 15, pp. 943–952, 2020.

74. T. Küstner, N. Fuin, K. Hammernik, A. Bustin, H. Qi, R. Hajhosseiny, P. G. Masci, R. Neji, D. Rueckert, R. M. Botnar *et al.*, "Cinenet: deep learning-based 3d cardiac cine mri reconstruction with multi-coil complex-valued 4d spatio-temporal convolutions," *Scientific Reports*, vol. 10, no. 1, pp. 1–13, 2020.

75. J. Zhang, A. Liu, M. Gao, X. Chen, X. Zhang, and X. Chen, "Ecg-based multi-class arrhythmia detection using spatio-temporal attention-based convolutional recurrent neural network," *Artificial Intelligence in Medicine*, vol. 106, p. 101856, 2020.

76. Q. Yao, R. Wang, X. Fan, J. Liu, and Y. Li, "Multi-class arrhythmia detection from 12-lead varied-length ecg using attention-based time-incremental convolutional neural network," *Information Fusion*, vol. 53, pp. 174–182, 2020.

77. G. Li, C. H. Lee, J. J. Jung, Y. C. Youn, and D. Camacho, "Deep learning for eeg data analytics: A survey," *Concurrency and Computation: Practice and Experience*, vol. 32, no. 18, p. e5199, 2020.

78. N. V. Saichand *et al.*, "Epileptic seizure detection using novel multilayer lstm discriminant network and dynamic mode koopman decomposition," *Biomedical Signal Processing and Control*, vol. 68, p. 102723, 2021.

79. M. Asadi-Aghbolaghi, A. Clapes, M. Bellantonio, H. J. Escalante, V. Ponce-López, X. Baró, I. Guyon, S. Kasaei, and S. Escalera, "A survey on deep learning based approaches for action and gesture recognition in image sequences," in *12th IEEE Intl. Conf. Automatic Face & Gesture Recog. (FG 2017)*, pp. 476–483.

80. P. Pareek and A. Thakkar, "A survey on video-based human action recognition: recent updates, datasets, challenges, and applications," *Artificial Intelligence Review*, vol. 54, no. 3, pp. 2259–2322, 2021.

81. Z. Du, S. Wu, D. Huang, W. Li, and Y. Wang, "Spatio-temporal encoder-decoder fully convolutional network for video-based dimensional emotion recognition," *IEEE Trans. Affective Computing*, vol. 12, no. 3, pp. 565–578, 2019.

82. G. Dong, K. G. Felker, A. Svyatkovskiy, W. Tang, and J. Kates-Harbeck, "Fully convolutional spatio-temporal models for representation learning in plasma science," *J. Machine Learning for Modeling and Computing*, vol. 2, no. 1, pp. 49–64, 2021.

83. K. Palaniappan, F. Bunyak, and S. S. Chaurasia, "Image analysis for ophthalmology: Segmentation and quantification of retinal vascular systems," in *Ocular Fluid Dynamics*. Springer, 2019, pp. 543–580.

84. K. Mittal and V. M. A. Rajam, "Computerized retinal image analysis-a survey," *Multimedia Tools and Applications*, vol. 79, pp. 22 389–22 421, 2020.

85. A. A. Abdulsahib, M. A. Mahmoud, M. A. Mohammed, H. H. Rasheed, S. A. Mostafa, and M. S. Maashi, "Comprehensive review of retinal blood vessel segmentation and classification techniques: Intelligent solutions for green computing in medical images,

current challenges, open issues, and knowledge gaps in fundus medical images," *Network Modeling Analysis in Health Informatics and Bioinformatics*, vol. 10, no. 1, pp. 1–32, 2021.

86. M. M. Fraz, P. Remagnino, A. Hoppe, B. Uyyanonvara, A. R. Rudnicka, C. G. Owen, and S. A. Barman, "Blood vessel segmentation methodologies in retinal images–a survey," *Computer Methods and Programs in Biomedicine*, vol. 108, no. 1, pp. 407–433, 2012.

87. N. Singh and L. Kaur, "A survey on blood vessel segmentation methods in retinal images," in *Int. Conf. Electronic Design, Computer Networks & Automated Verification (EDCAV)*. IEEE, 2015, pp. 23–28.

88. M. Niemeijer, J. Staal, B. van Ginneken, M. Loog, and M. D. Abramoff, "Comparative study of retinal vessel segmentation methods on a new publicly available database," in *Medical Imaging 2004: Image Processing*, vol. 5370. International Society for Optics and Photonics, 2004, pp. 648–656.

89. A. Hoover, V. Kouznetsova, and M. Goldbaum, "Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response," *IEEE Trans. Medical Imaging*, vol. 19, no. 3, pp. 203–210, 2000.

90. R. Kälviäinen and H. Uusitalo, "Diaretdb1 diabetic retinopathy database and evaluation protocol," in *Medical Image Understanding and Analysis*, vol. 2007. Citeseer, 2007, p. 61.

91. B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 2921–2929.

92. S. Desai and H. G. Ramaswamy, "Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization," in *Proc. IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 983–991.

93. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Intl. Conf. Computer Vision*, 2017, pp. 618–626.

94. A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *IEEE Winter Conf. Applications of Computer Vision (WACV)*, 2018, pp. 839–847.

95. M. Zhang, Y. Zhou, J. Zhao, Y. Man, B. Liu, and R. Yao, "A survey of semi-and weakly supervised semantic segmentation of images," *Artificial Intelligence Review*, vol. 53, no. 6, pp. 4259–4288, 2020.

96. F. Shao, L. Chen, J. Shao, W. Ji, S. Xiao, L. Ye, Y. Zhuang, and J. Xiao, "Deep learning for weakly-supervised object detection and localization: A survey," *Neurocomputing*, vol. 496, pp. 192–207, 2022.

97. Z.-H. Zhou, "A brief introduction to weakly supervised learning," *National Science Review*, vol. 5, no. 1, pp. 44–53, 2018.

98. D. Zhang, J. Han, G. Cheng, and M. Yang, "Weakly supervised object localization and detection: A survey," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 5866–5885, 2021.

99. B. Ben-David and M. Icht, "Oral-diadochokinetic rates for Hebrew-speaking healthy ageing population: non-word versus real-word repetition," *Int. Journal of Language & Communication Disorders*, vol. 52, no. 3, pp. 301–310, 2017.

100. M. Duranovic and S. Sehic, "The speed of articulatory movements involved in speech production in children with dyslexia," *Journal of Learning Disabilities*, vol. 46, no. 3, pp. 278–286, 2013.

101. S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

102. F. Bunyak, N. Shiraishi, K. Palaniappan, T. Lever, L. Avivi-Arber, and K. Takahashi, "Development of semi-automatic procedure for detection and tracking of fiducial markers for orofacial kinematics during natural feeding," in *IEEE Engineering in Medicine and Biology Society (EMBC)*, 2017, pp. 580–583.

103. L. Welby, H. Caudill, G. Yitsege, A. Hamad, F. Bunyak, I. Zohn, T. Maynard, A. LaMantia, D. Mendelowitz, and T. Lever, "Persistent feeding and swallowing deficits in a mouse model of 22q11. 2 deletion syndrome," *Frontiers in Neurology*, vol. 11, p. 4, 2020.

104. R. M. Haralick and L. G. Shapiro, *Computer and robot vision*, ser. Computer and Robot Vision. Addison-Wesley Pub. Co., 1993.

105. M. Slaney *et al.*, "An efficient implementation of the patterson-holdsworth auditory filter bank," *Apple Computer, Perception Group, Tech. Rep*, vol. 35, no. 8, 1993.

106. Y. Fu and T. Huang, "Human age estimation with regression on discriminative aging manifold," *IEEE Trans. Multimedia*, vol. 10, no. 4, pp. 578–584, 2008.

107. G. Guo, Y. Fu, C. Dyer, and T. Huang, "Image-based human age estimation by manifold learning and locally adjusted robust regression," *IEEE Trans. Image Processing*, vol. 17, no. 7, pp. 1178–1188, 2008.

108. C. T. Sasaki and E. M. Weaver, "Physiology of the larynx," *The American journal of medicine*, vol. 103, no. 5, pp. 9S–18S, 1997.

109. J. Dankbaar and F. Pameijer, "Vocal cord paralysis: anatomy, imaging and pathology," *Insights into imaging*, vol. 5, no. 6, pp. 743–751, 2014.

110. M. Weinberger and D. Doshi, "Vocal cord dysfunction: a functional cause of respiratory distress," *Breathe*, vol. 13, no. 1, pp. 15–21, 2017.

111. A. Rajaei, E. Barzegar Bafrooei, F. Mojiri, and M. H. Nilforoush, "The occurrence of laryngeal penetration and aspiration in patients with glottal closure insufficiency," *ISRN Otolaryngology*, vol. 2014, 2014.

112. S. J. S. Toutounchi, M. Eydi, S. E. Golzari, M. R. Ghaffari, and N. Parvizian, "Vocal cord paralysis and its etiologies: a prospective study," *J. Cardiovascular and Thoracic Research*, vol. 6, no. 1, p. 47, 2014.

113. X. Hu, S. Y. Eunhee, and J. H. Ryu, "Aspiration-related deaths in 57 consecutive patients: autopsy study," *PLoS ONE*, vol. 9, no. 7, 2014.

114. T. Kim, K. Goodhart, J. E. Aviv, R. L. Sacco, B. Diamond, S. Kaplan, and L. G. Close, "Feesst: a new bedside endoscopic test of the motor and sensory components of swallowing," *Annals of Otology, Rhinology & Laryngology*, vol. 107, no. 5, pp. 378–387, 1998.

115. L. A. Shock, B. C. Gallemore, C. J. Hinkel, M. M. Szewczyk, B. L. Hopewell, M. J. Allen, L. A. Thombs, and T. E. Lever, "Improving the utility of laryngeal adductor reflex testing: A translational tale of mice and men," *Otolaryngology–Head and Neck Surgery*, vol. 153, no. 1, pp. 94–101, 2015.

116. T. E. Lever, A. M. Kloepper, I. Deninger, A. Hamad, B. L. Hopewell, A. K. Ovaitt, M. Szewczyk, F. Bunyak, B. Zitsch, B. Blake *et al.*, "Advancing laryngeal adductor reflex testing beyond sensory threshold detection," *Dysphagia*, pp. 1–21, 2021.

117. M. M. Haney, A. Hamad, H. G. Woldu, M. Ciucci, N. Nichols, F. Bunyak, and T. E. Lever, "Recurrent laryngeal nerve transection in mice results in translational upper airway dysfunction," *J. Comparative Neurology*, vol. 528, no. 4, pp. 574–596, 2020.

118. J. Zhang, C. Li, M. M. Rahaman, Y. Yao, P. Ma, J. Zhang, X. Zhao, T. Jiang, and M. Grzegorzek, "A comprehensive review of image analysis methods for microorganism counting: from classical image processing to deep learning approaches," *Artificial Intelligence Review*, pp. 1–70, 2021.

119. T. G. Debelee, F. Schwenker, A. Ibenthal, and D. Yohannes, "Survey of deep learning in breast cancer image analysis," *Evolving Systems*, vol. 11, no. 1, pp. 143–163, 2020.

120. X. Liu, K. Li, R. Yang, and L. Geng, "Review of deep learning based automatic segmentation for lung cancer radiotherapy," *Frontiers in Oncology*, vol. 11, p. 2599, 2021.

121. M. G. Bandyk, D. R. Gopireddy, C. Lall, K. Balaji, and J. Dolz, "Mri and ct bladder segmentation from classical to deep learning based approaches: Current limitations and lessons," *Computers in Biology and Medicine*, p. 104472, 2021.

122. S. Bhattacharya, P. K. R. Maddikunta, Q. Pham, T. R. Gadekallu, C. L. Chowdhary, M. Alazab, M. J. Piran *et al.*, "Deep learning and medical image processing for coronavirus (covid-19) pandemic: A survey," *Sustainable Cities and Society*, vol. 65, p. 102589, 2021.

123. R. Girshick, "Fast r-cnn," in *Proc. IEEE Conf. Comp. Vision*, 2015, pp. 1440–1448.

124. K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proc. of IEEE Conf. Comp. Vision*, 2017, pp. 2961–2969.

125. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE Conf. Comp. Vision and Pattern Recog.*, 2009, pp. 248–255.

126. K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. European Conf. Comp. Vision*. Springer, 2016, pp. 630–645.

127. Y. Wang, A. Hamad, T. Lever, and F. Bunyak, "Orthogonal region selection network for laryngeal closure detection in laryngoscopy videos," in *Annual Int. Conf. IEEE Eng. in Medicine & Biology Society*, 2020, pp. 2167–2172.

128. R. C. Gonzalez, R. E. Woods, and B. R. Masters, "Digital image processing," 2009.

129. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comp. Vision and Pattern Recog.*, 2016, pp. 770–778.

130. T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Comput.*, vol. 10, no. 7, pp. 1895–1923, 1998.

131. A. Gupta and A. Periakaruppan, "Intracranial dural arteriovenous fistulas: a review," *The Indian Journal of Radiology & Imaging*, vol. 19, no. 1, p. 43, 2009.

132. M. S. Aboian, D. J. Daniels, S. K. Rammos, E. Pozzati, and G. Lanzino, "The putative role of the venous system in the genesis of vascular malformations," *Neurosurgical Focus*, vol. 27, no. 5, p. E9, 2009.

133. O. V. Glinskii, V. H. Huxley, V. V. Glinskii, L. J. Rubin, and V. V. Glinsky, "Pulsed estrogen therapy prevents post-ovx porcine dura mater microvascular network weakening via a pdgf-bb-dependent mechanism," *PLoS ONE*, vol. 8, no. 12, p. e82900, 2013.

134. T. Wilson, "Resolution and optical sectioning in the confocal microscope," *Journal of Microscopy*, vol. 244, no. 2, pp. 113–121, 2011.

135. S. Moccia, E. De Momi, S. El Hadji, and L. S. Mattos, "Blood vessel segmentation algorithms—review of methods, datasets and evaluation metrics," *Computer Methods and Programs in Biomedicine*, vol. 158, pp. 71–91, 2018.

136. F. Zhao, Y. Chen, Y. Hou, and X. He, "Segmentation of blood vessels using rule-based and machine-learning-based methods: a review," *Multimedia Systems*, vol. 25, no. 2, pp. 109–118, 2019.

137. Y. M. Kassim, R. J. Maude, and K. Palaniappan, "Sensitivity of cross-trained deep cnns for retinal vessel extraction," in *Annual Intl. Conf. IEEE Engineering in Medicine and Biology Society (EMBC)*, 2018, pp. 2736–2739.

138. Y. M. Kassim and K. Palaniappan, "Extracting retinal vascular networks using deep learning architecture," in *IEEE Intel. Conf. Bioinformatics and Biomedicine (BIBM)*, 2017, pp. 1170–1174.

139. M. M. H. Shuvo, Y. M. Kassim, F. Bunyak, O. V. Glinskii, L. Xie, V. V. Glinsky, V. H. Huxley, M. M. Thakkar, and K. Palaniappan, "Multi-focus image fusion for confocal microscopy using u-net regression map," in *IEEE Intl. Conf. Pattern Recognition (ICPR)*, 2021, pp. 4317–4323.

140. S. M. Pizer, E. P. Amburn, J. D. Austin, R. Cromartie, A. Geselowitz, T. Greer, B. ter Haar Romeny, J. B. Zimmerman, and K. Zuiderveld, "Adaptive histogram equalization and its variations," *Computer Vision, Graphics, and Image Processing*, vol. 39, no. 3, pp. 355–368, 1987.

141. Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2018, pp. 3–11.

142. K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT Press, 2012.

143. I. D. Mayergoyz, *Mathematical models of hysteresis and their applications*. Academic Press, 2003.

144. O. V. Glinskii, V. H. Huxley, L. Xie, F. Bunyak, K. Palaniappan, and V. V. Glinsky, "Complex non-sinus-associated pachymeningeal lymphatic structures: interrelationship with blood microvasculature," *Frontiers in Physiology*, vol. 10, p. 1364, 2019.

145. J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

146. L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

147. Y. Ouassit, S. Ardchir, M. Yassine E. G., and M. Azouazi, "A brief survey on weakly supervised semantic segmentation." *Intl. J. Online & Biomedical Engineering*, vol. 18, no. 10, 2022.

148. H. Ding, Q. Cen, X. Si, Z. Pan, and X. Chen, "Automatic glottis segmentation for laryngeal endoscopic images based on u-net," *Biomedical Signal Processing and Control*, vol. 71, p. 103116, 2022.

149. M. Pedersen, C. F. Larsen, B. Madsen, and M. Eeg, "Localization and quantification of glottal gaps on deep learning segmentation of vocal folds," *Scientific Reports*, vol. 13, no. 1, p. 878, 2023.

150. A. Bandini, S. Smaoui, and C. M. Steele, "Automated pharyngeal phase detection and bolus localization in videofluoroscopic swallowing study: Killing two birds with one stone?" *Computer Methods and Programs in Biomedicine*, vol. 225, p. 107058, 2022.

151. M. Früh, M. Fischer, A. Schilling, S. Gatidis, and T. Hepp, "Weakly supervised segmentation of tumor lesions in pet-ct hybrid imaging," *J. Medical Imaging*, vol. 8, no. 5, pp. 054 003–054 003, 2021.

152. B. Kopczynski, E. Niebudek-Bogusz, W. Pietruszewska, and P. Strumillo, "Segmentation of glottal images from high-speed videoendoscopy optimized by synchronous acoustic recordings," *Sensors*, vol. 22, no. 5, p. 1751, 2022.

153. A. M. Kist, S. Dürr, A. Schützenberger, and M. Döllinger, "Openhsv: An open platform for laryngeal high-speed videoendoscopy," *Scientific Reports*, vol. 11, no. 1, 2021.

154. A. S. Hamad, Y. Y. Wang, T. E. Lever, and F. Bunyak, "Ensemble of deep cascades for detection of laryngeal adductor reflex events in endoscopy videos," *2020 IEEE International Conference on Image Processing (ICIP)*.

155. J. Xue, L. Yin, Z. Lan, M. Long, G. Li, Z. Wang, and X. Xie, "3d dct based image compression method for the medical endoscopic application," *Sensors*, vol. 21, no. 5, p. 1817, 2021.

156. X. Zhang, X. Chen, L. Yao, C. Ge, and M. Dong, "Deep neural network hyperparameter optimization with orthogonal array tuning," in *Neural Information Processing: 26th Intl. Conf., ICONIP 2019*. Springer, pp. 287–295.

157. A. Darwish, D. Ezzat, and A. E. Hassanien, "An optimized model based on convolutional neural networks and orthogonal learning particle swarm optimization algorithm for plant diseases diagnosis," *Swarm and Wvolutionary Computation*, vol. 52, p. 100616, 2020.

158. J. Brownlee, "A gentle introduction to the rectified linear unit (relu)," *Machine Learning Mastery*, vol. 6, 2019.

159. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

160. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.

161. M. B. Muhammad and M. Yeasin, "Eigen-cam: Class activation map using principal components," in *IEEE Intl. Joint Conf. Neural Networks*, 2020, pp. 1–7.

162. P. Jiang, C. Zhang, Q. Hou, M. Cheng, and Y. Wei, "Layercam: Exploring hierarchical class activation maps for localization," *IEEE Trans. Image Processing*, vol. 30, pp. 5875–5888, 2021.

163. A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

164. F. Karlsson, E. Schalling, K. Laakso, K. Johansson, and L. Hartelius, "Assessment of speech impairment in patients with parkinson's disease from acoustic quantifications of oral diadochokinetic sequences," *J. the Acoustical Society of America*, vol. 147, no. 2, pp. 839–851, 2020.

165. S. Hahm and J. Wang, "Parkinson's condition estimation using speech acoustic and inversely mapped articulatory data," 2015.

166. T. Grósz, R. Busa-Fekete, G. Gosztolya, and L. Tóth, "Assessing the degree of nativeness and parkinson's condition using gaussian processes and deep rectifier neural networks," 2015.

167. X. Ma, H. Yang, Q. Chen, D. Huang, and Y. Wang, "Depaudionet: An efficient deep model for audio based depression classification," in *Proc. the 6th Int. Workshop on Audio/Visual Emotion Challenge*, 2016, pp. 35–42.

168. V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

169. A. M. Saxe, J. L. McClelland, and S. Ganguli, "Exact solutions to the nonlinear dynamics of learning in deep linear neural networks," *arXiv preprint arXiv:1312.6120*, 2013.

170. D. Xie, J. Xiong, and S. Pu, "All you need is beyond a good init: Exploring better solution for training extremely deep convolutional neural networks with orthonormality and modulation," in *Proc. IEEE Conf. on Comp. Vision and Pattern Recog.*, 2017, pp. 6176–6185.

171. T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Conf. Comp Vision*, 2017, pp. 2980–2988.

172. J. Lohscheller, H. Toy, F. Rosanowski, U. Eysholdt, and M. Döllinger, "Clinically evaluated procedure for the reconstruction of vocal fold vibrations from endoscopic digital high-speed videos," *Medical image analysis*, vol. 11, no. 4, pp. 400–413, 2007.

173. X. Qin, S. Wang, and M. Wan, "Improving reliability and accuracy of vibration parameters of vocal folds based on high-speed video and electroglottography," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 6, pp. 1744–1754, 2009.

174. A. Hamad, I. Ersoy, and F. Bunyak, "Improving nuclei classification performance in H&E stained tissue images using fully convolutional regression network and convolutional neural network," in *IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, 2018, pp. 1–6.

175. R. Bao, N. M. Al-Shakarji, F. Bunyak, and K. Palaniappan, "DMNet: dual-stream marker guided deep network for dense cell segmentation and lineage tracking," in *Proc. IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3361–3370.

176. F. Xing, Y. Xie, H. Su, F. Liu, and L. Yang, "Deep learning in microscopy image analysis: A survey," *IEEE Trans. Neural Networks and Learning Systems*, vol. 29, no. 10, pp. 4550–4568, 2017.

177. S. A. Bala and S. Kant, "Deep learning-based model architectures for cardiac mri segmentation: A survey," *Intl. J. Innovative Science, Engineering & Technology*, pp. 129–135, 2020.

178. J. Liu, Y. Pan, M. Li, Z. Chen, L. Tang, C. Lu, and J. Wang, "Applications of deep learning to mri images: A survey," *Big Data Mining and Analytics*, vol. 1, no. 1, pp. 1–18, 2018.

179. R. Pelapur, V. S. Prasath, F. Bunyak, O. V. Glinskii, V. V. Glinsky, V. H. Huxley, and K. Palaniappan, "Multi-focus image fusion using epifluorescence microscopy for robust vascular segmentation," in *36th Annual International Conf. the IEEE Engineering in Medicine and Biology Society*, 2014, pp. 4735–4738.

180. A. Raphael, Z. Dubinsky, D. Iluz, and N. S. Netanyahu, "Neural network recognition of marine benthos and corals," *Diversity*, vol. 12, no. 1, p. 29, 2020.

181. F. Bu and D. E. Chang, "Feedback gradient descent: Efficient and stable optimization with orthogonality for dnns," in *Proc. AAAI Conference on Artificial Intelligence*, vol. 36, no. 6, 2022, pp. 6106–6114.

182. Z. Zhang, W. Ma, Y. Wu, and G. Wang, "Self-orthogonality module: A network architecture plug-in for learning orthogonal filters," in *Proc. IEEE/CVF Winter Conf. Applications of Computer Vision*, 2020, pp. 1050–1059.

183. Y. Wang, B. Lei, A. Elazab, E. Tan, W. Wang, F. Huang, X. Gong, and T. Wang, "Breast cancer image classification via multi-network features and dual-network orthogonal low-rank learning," *IEEE Access*, vol. 8, pp. 27 779–27 792, 2020.

184. H.-C. Chen, Y.-M. Jen, C.-H. Wang, J.-C. Lee, and Y.-S. Lin, "Etiology of vocal cord paralysis," *ORL*, vol. 69, no. 3, pp. 167–171, 2007.

185. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comp. Vision and Pattern Recog.*, 2016, pp. 779–788.

186. M. Roig-Quilis and A. Rodríguez-Palmero, "Oromotor disorders in a paediatric neurology unit. their classification and clinical course," *Revista de Neurologia*, vol. 47, no. 10, pp. 509–516, 2008.

187. M. Roig-Quilis, "Oromotor dysfunction in neuromuscular disorders: evaluation and treatment," in *Neuromuscular Disorders of Infancy, Childhood, and Adolescence*. Elsevier, 2015, pp. 958–975.

188. M. M. van der Graaff, W. Grolman, E. J. Westermann, H. C. Boogaardt, H. Koelman, A. J. van der Kooi, M. A. Tijssen, and M. de Visser, "Vocal cord dysfunction in amyotrophic lateral sclerosis: four cases and a review of the literature," *Archives of Neurology*, vol. 66, no. 11, pp. 1329–1333, 2009.

189. R. D. Kent, "Nonspeech oral movements and oral motor disorders: A narrative review," *American Journal of Speech-Language Pathology*, vol. 24, no. 4, pp. 763–789, 2015.

190. S. P. das Neves, N. Delivanoglou, and S. Da Mesquita, "Cns-draining meningeal lymphatic vasculature: Roles, conundrums and future challenges," *Frontiers in Pharmacology*, vol. 12, p. 964, 2021.

191. W. H. Organization, *Neurological disorders: public health challenges*. World Health Organization, 2006.

192. S. M. Anwar, M. Majid, A. Qayyum, M. Awais, M. Alnowami, and M. K. Khan, "Medical image analysis using convolutional neural networks: a review," *J. Medical Systems*, vol. 42, no. 11, pp. 1–13, 2018.

193. G. Mohan and M. M. Subashini, "Mri based medical image analysis: Survey on brain tumor grade classification," *Biomedical Signal Processing and Control*, vol. 39, pp. 139–161, 2018.

194. A. Mansoor, U. Bagci, B. Foster, Z. Xu, G. Z. Papadakis, L. R. Folio, J. K. Udupa, and D. J. Mollura, "Segmentation and image analysis of abnormal lungs at ct: current approaches, challenges, and future trends," *Radiographics*, vol. 35, no. 4, pp. 1056–1076, 2015.

195. J. E. Aviv, S. T. Kaplan, J. E. Thomson, J. Spitzer, B. Diamond, and L. G. Close, "The safety of flexible endoscopic evaluation of swallowing with sensory testing (feesst): an analysis of 500 consecutive evaluations," *Dysphagia*, vol. 15, no. 1, pp. 39–44, 2000.

196. E. Moen, D. Bannon, T. Kudo, W. Graf, M. Covert, and D. Van Valen, "Deep learning for cellular image analysis," *Nature Methods*, pp. 1–14, 2019.

197. J. Ma, Y. Song, X. Tian, Y. Hua, R. Zhang, and J. Wu, "Survey on deep learning for pulmonary medical imaging," *Frontiers of Medicine*, pp. 1–20, 2019.

198. M. A. Mazurowski, M. Buda, A. Saha, and M. R. Bashir, "Deep learning in radiology: An overview of the concepts and a survey of the state of the art with focus on mri," *J. Magnetic Resonance Imaging*, vol. 49, no. 4, pp. 939–954, 2019.

199. K. Palaniappan, F. Bunyak, and S. Chaurasia, "Image analysis for ophthalmology: Segmentation and quantification of retinal vascular systems," in *Ocular Fluid Dynamics*. Springer, 2019, pp. 543–580.

200. S. Saadatnejad, M. Oveisi, and M. Hashemi, "Lstm-based ecg classification for continuous monitoring on personal wearable devices," *J. Biomedical and Health Informatics*, vol. 24, no. 2, pp. 515–523, 2019.

201. D. Voigt, M. Döllinger, A. Yang, U. Eysholdt, and J. Lohscheller, "Automatic diagnosis of vocal fold paresis by employing phonovibrogram features and machine learning methods," *Computer Methods and Programs in Biomedicine*, vol. 99, no. 3, pp. 275–288, 2010.

202. A. Verikas, A. Gelzinis, M. Bacauskiene, M. Hållander, V. Uloza, and M. Kaseta, "Combining image, voice, and the patient's questionnaire data to categorize laryngeal disorders," *Artificial Intelligence in Medicine*, vol. 49, no. 1, pp. 43–50, 2010.

203. R. Schwarz, U. Hoppe, M. Schuster, T. Wurzbacher, U. Eysholdt, and J. Lohscheller, "Classification of unilateral vocal fold paralysis by endoscopic digital high-speed recordings and inversion of a biomechanical model," *IEEE Trans. Biomed. Engr.*, vol. 53, no. 6, pp. 1099–1108, 2006.

204. I. Miliaresi, K. Poutos, and A. Pikrakis, "Combining acoustic features and medical data in deep learning networks for voice pathology classification," in *28th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021, pp. 1190–1194.

205. S.-H. Fang, Y. Tsao, M.-J. Hsiao, J.-Y. Chen, Y.-H. Lai, F.-C. Lin, and C.-T. Wang, "Detection of pathological voice using cepstrum vectors: A deep learning approach," *J. Voice*, vol. 33, no. 5, pp. 634–641, 2019.

206. H. Guan and A. Lerch, "Learning strategies for voice disorder detection," in *13th Intl. Conf. Semantic Computing (ICSC)*. IEEE, 2019, pp. 295–301.

207. S. P. Singh, L. Wang, S. Gupta, H. Goli, P. Padmanabhan, and B. Gulyás, "3d deep learning on medical images: a review," *Sensors*, vol. 20, no. 18, p. 5097, 2020.

208. A. S. Lundervold and A. Lundervold, "An overview of deep learning in medical imaging focusing on mri," *Zeitschrift für Medizinische Physik*, vol. 29, no. 2, pp. 102–127, 2019.

209. E. W. Forgy, "Cluster analysis of multivariate data: efficiency versus interpretability of classifications," *Biometrics*, vol. 21, pp. 768–769, 1965.

210. G. J. McLachlan, S. X. Lee, and S. I. Rathnayake, "Finite mixture models," *Annual review of statistics and its application*, vol. 6, pp. 355–378, 2019.

211. G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

212. E. A. Nadaraya, "On estimating regression," *Theory of Probability & Its Applications*, vol. 9, no. 1, pp. 141–142, 1964.

213. G. S. Watson, "Smooth regression analysis," *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 359–372, 1964.

214. A. Das and P. Rad, "Opportunities and challenges in explainable artificial intelligence (xai): A survey," *arXiv preprint arXiv:2006.11371*, 2020.

VITA

Yangyang Wang is a Doctor of Philosophy in Computer Science. She received her Bachelor's degree from the University of Electronic Science and Technology of China. Then she received her Master's degree and Ph.D. in Computer Science at the University of Missouri-Columbia in the United States.

Her research topic focuses on deep learning, image processing, signal processing, and biomedical informatics.