INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

LungRN+NL: An Improved Adventitious Lung Sound Classification Using non-local block ResNet Neural Network with Mixup Data Augmentation

Yi Ma, Xinzi Xu, and Yongfu Li

Department of Micro-Nano Electronics and MoE Key Lab of Artificial Intelligence
Shanghai Jiao Tong University, Shanghai, China
yongfu.li@sjtu.edu.cn

Abstract

Performing an automated adventitious lung sound detection is a challenging task since the sound is susceptible to noises (heartbeat, motion artifacts, and audio sound) and there is subtle discrimination among different categories. An adventitious lung sound classification model, LungRN+NL, is proposed in this work, which has demonstrated a drastic improvement compared to our previous work and the state-of-the-art models. This new model has incorporated the non-local block in the ResNet architecture. To address the imbalance problem and to improve the robustness of the model, we have also incorporated the mixup method to augment the training dataset. Our model has been implemented and compared with the state-of-the-art works using the official ICBHI 2017 challenge dataset and their evaluation method. As a result, LungRN+NL has achieved a performance score of 52.26%, which is improved by 2.1-12.7% compared to the state-of-the-art models.

Index Terms: adventitious lung sounds classification, mixup, data augmentation, convolutional neural network, non-local block

1. Introduction

Respiratory-related diseases, such as pneumonia and asthma, are one of the leading causes of death in the world [1]. Auscultation plays an important role in early diagnosis of these diseases [2]. For example, crackle (the earliest sign of idiopathic pulmonary fibrosis [2]) and wheeze (asthma and chronic obstructive lung disease [2]) can be recorded through a stethoscope and analyzed by a professional medical practitioner. Despite the ease of testing method, in the case of sudden, massive infectious respiratory-related diseases outbreak, such as the pneumonia complication from the coronavirus [3], a shortage of medical practitioners can further exacerbate the spread of the diseases and increase of death rate. According to the World Health Organization statistics, 45% of the state members reported having less than 1 physician per 1000 population [4]. Therefore, it is necessary to develop an automatic respiratory detection to reduce the workload for medical practitioners.

In 2017, the International Conference on Biomedical and Health Informatics (ICBHI) has organized a classification competition based on the adventitious lung sounds dataset for the research community [5]. This dataset has inspired several research groups to explore the problem with the statistical method (HMM-GMM) and machine learning methods (boosted decisional tree and SVM) and classify the dataset into 3 types of adventitious lung sounds [6–8]. The introduction of deep learning has greatly improved detection accuracy. For example, [9] proposed to use a deep recurrent neural network with a noise-masking model, and [10] proposed to use a hybrid CNN-RNN model to perform classification. In [11, 12], we have proposed a LungRBN model, which uses short-time Fourier transform (STFT) and wavelet feature extraction methods together with a product of two ResNet models through a fully connected layer to achieve the best state-of-the-art accuracy. However, less attention has been paid to finding ways to automatically augment existing data to achieve a significant breakthrough in detection accuracy.

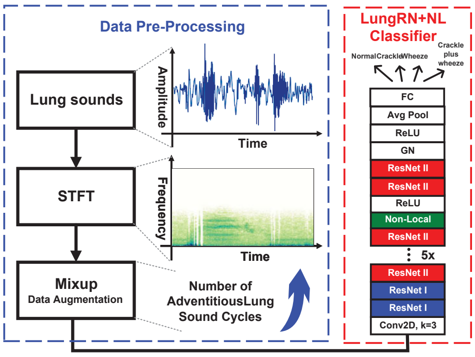

Figure 1: *An conceptual idea of the LungRN+NL with mixup data augmentation.*

To overcome this challenge, we propose an improved adventitious Lung Sound Classification, LungRN+NL, incorporate a non-local layer in ResNet neural network with a mixup data augmentation method. Considering the key discrimination among different categories, we choose short-time Fourier transform (STFT), a time-frequency analysis method, to extract features from lung sounds. The key contributions of our works are summarized as follows:

1. Based on the medical characteristic of adventitious lung sound ICBHI2017 dataset [5], we propose to use mixup [13] to augment a new dataset for our LungRN+NL model.
2. We introduce a non-local block [14, 15] in the ResNet architecture to calculates the relationship across a different position in the STFT spectrogram across time and frequency domain.
3. We verify our proposed LungRN+NL model on the ICBHI2017 dataset and we are able to achieve a performance score of 52.26%, which is an improvement of 2.1-12.7% compared to the state-of-the-art modes.

The remainder of the paper is organized as follows. Section 2 describes the data augmentation used for the adventitious lung sounds classification. Section 2 details the non-local module and our proposed network structure. Section 4 presents the

http://dx.doi.org/10.21437/Interspeech.2020-2487

Table 1: *Distribution of Classes in ICBHI 2017 and after augmentation method.*

	Before augmentation		**After augmentation**	
	Amount	Ratio	Amount[1]	Ratio[2]
Normal	3643	0.528	2073	0.242
Crackle [C]	1874	0.271	3585	0.419
Wheeze [W]	886	0.128	1470	0.172
C+W	495	0.072	1431	0.167
Total	6898	1	8559	1

[1] The amount is number of samples used for training.

[2] Average categories ratio after augmentation. It was computed in an epoch.

experimental methods and results with discussions. Finally, the conclusion and summary are given in Section 5.

2. ICBHI 2017 Dataset and Data Augmentation

2.1. Pre-processing ICBHI 2017 Dataset

ICBHI 2017 Dataset: In 2017, International Conference on Biomedical and Health Informatics (ICBHI) has officially released the first open-source adventitious lung sound dataset, which consists of a total of 5.5-hour recordings, containing annotated respiratory cycles from 126 subjects [5]. A record is defined as the lung sounds collected from one patient and a cycle is defined as an adventitious lung sound classification label. Hence, this dataset comprises 3,642 “normal”, 1864 “crackle”, 886 “wheeze”, and 506 “crackle plus wheeze” cycles with a total of 6,898 cycles. This dataset was collected from several hospitals using different voice recording equipment. Thus, the sampling frequency, background noise level, and the number of respiratory cycles vary among patients.

Pre-processing: Since lung sound is feeble and susceptible to different environment noises [2], such as heartbeat, motion artifacts and audio sound, a 5-th order Butterworth band-pass filter helps to retain the frequency of interest from 100 to 2,000Hz. Short-time Fourier transform (STFT) method [16] extracts important adventitious lung sound features in time and frequency domain for each respiratory cycle. Considering the nonlinear and non-stationary characteristic of adventitious lung sound [2], we used a Hanning window of 0.02 seconds window length and 0.01 seconds in hop length between two adjacent windows. All recordings are processed with the min-max normalization method [17] to standardize the data across different recording devices. Since each respiratory cycle has varying lengths, we have used stretching and compressing methods to fix the size of each STFT spectrogram record with 128×128 matrices.

2.2. “Mixup” Data Augmentation method

Imbalance dataset is practically a common problem in all the medical classification tasks due to the low probability of obtaining abnormal samples [18]. As shown in Table 1, the ratio between the “normal” and “crackle plus wheeze” classes is 0.072 or 7.36× different. Data augmentation method is necessary to address the imbalanced problem [19].

Since adventitious lung sound is feeble and susceptible to different environment noises [2], traditional data augmentation methods [20] generate new samples by changing the characteristic of the signals in time or frequency domain. However, these methods is not effective in improvement in detection accuracy, which is observed through our experimental result in Section 4.2.

To address the aforementioned problem, we have proposed to mixup method [13], which is a data-agnostic data augmentation method that makes decision boundaries transit linearly from class to class and provides a smoother estimate of uncertainty. To apply the mixup method in the ICBHI 2017 dataset, we first have to understand the characteristics of each adventitious lung sound type. For example, the characteristics of crackle sound tends to have short and explosive sound, which can be viewed as a specific sound event with normal lung sound as background. Hence, it is reasonable to combine “normal” cycles with “crackle” cycles to increase the number of “crackle” cycles in the dataset. Similarly, the number of “wheeze” cycles in the dataset can be increased by combining the “normal” cycles and “wheeze” cycles. The new “crackle plus wheeze” cycles are obtained by mixing the “crackle” and “wheeze” cycles.

To combine two randomly selected samples with their labels in the training dataset and generate a new set of sample and label, which can be described as follow:

$$\begin{aligned} \widetilde{x} &= \lambda x_i + (1-\lambda)x_j, \\ \widetilde{y} &= \lambda y_i + (1-\lambda)y_j, \end{aligned} \tag{1}$$

where (x_i, x_j) are two feature vectors and (y_i, y_j) are one-hot encoded class labels of the two features. $\lambda \in [0, 1]$ is a random number generated according to Beta distribution [17].

Although one-hot encoded label vector is popular in classification task, it might not be appropriate for adventitious lung sound classification project. In reality, there is no quantitative standard to diagnosis adventitious lung sounds and medical staff analyzes each sample according to their experience. Furthermore, the severity of the illness varies among patients and it is hard to obtain the identical results for the same type of illness. Therefore, to address the aforementioned problem, we use probability function to represent the generated label vector with the value varying from 0 to 1, which is determined by λ.

Figure 2 illustrates the procedure of using a mixup method for the adventitious lung sounds dataset used in our work. For each batch of data, the “normal” cycles are combined with the “crackle” and “wheeze” cycles using the mixup method to generate a new set of “crackle” and “wheeze” cycles, respectively. The “crackle” and “wheeze” cycles are combined to generate “crackle plus wheeze” samples. The new label vectors generated by the mixup method are computed by 1. Note that for the labels newly generated for “crackle plus wheeze” cycles are kept in a one-hot encoding format.

3. LungRN+NL Neural Network Architecture

Non-Local Block: Compared to our previous works [11, 12], we have introduced a non-local block [14, 15] in the ResNet architecture to calculates the relationship across a different position in STFT spectrogram across time and frequency domain. The non-local layer is defined as:

$$y_i = \frac{1}{C(x)} \sum_{\forall j} f(x_i, x_j) g(x_j), \tag{2}$$

where y_i is the output of the non-local layer and it is computed by combining x_i and all possible positions x_j in the input spectrogram. The relationship between x_i and x_j is computed by

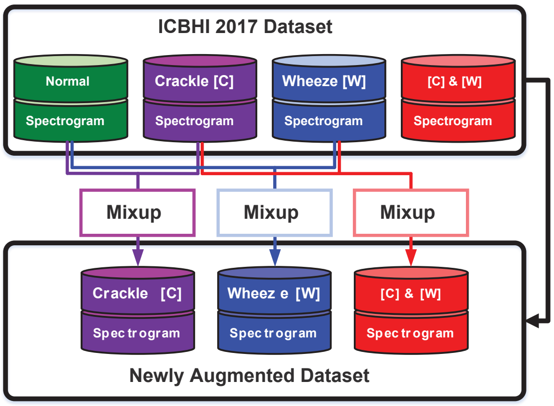

Figure 2: *The proposed data augmentation result using the mixup method with ICBHI 2017 dataset.*

a pairwise function f. The function g represents the response of input signal at position j and it is implemented with a convolution operation with a kernel size of 1. $\sum_{\forall j} f(x_i, x_j)$ is normalized by the function C.

Ref. [14] provides several different methods of implementing the function f, including dot product, concatenation, embedded Gaussian. After our preliminary study on our LungRN+NL model with the ICBHI 2017 dataset, we choose to use the embedded Gaussian function to compute the relationship between x_i and x_j. The embedded Gaussian function is given as follow:

$$f(x_i, x_j) = e^{\theta(x_i)^T \phi(x_j)}, \tag{3}$$

where θ and ϕ are two convolution operations with a kernel size of 1. Note that the embedded Gaussian function has a self-attention mechanism, which helps to further reduce the loss [21].

LungRN+NL Neural Network: Figure 3 illustrates our proposed neural network architecture with the STFT spectrogram as inputs for adventitious lung sounds classification. To enlarge the receptive field [22] when the signals propagate through the layers in the neural network, we propose to use a 2D convolutional layer to expands the number of time and frequency channels Subsequently, we use ResNet-I layers to reduce the size of STFT spectrogram.

To learn the characteristic of lung sounds through time and frequency domain, we employ several ResNet-II layers with the 2D convolution to process spectrogram. A non-local layer is inserted between ResNet-II layers to break the local time and frequency limit from the convolutional neural network, which means the non-local layer computes the relationship of a position with other positions across time and frequency domain. Lastly, after the signals propagated through the ResNet-II, we have used a group normalization GN [17], nonlinear ReLU activation function [17] and average pooling layer [17] with a fully-connection layer to classify the cycles into 4 different classes.

The details of the ResNet-I, ResNet-II and non-local layers are shown in Figure 3, respectively. The characteristic *k* means the kernel size of the convolutional operation and in non-local part, *H*, *W*, *C* represents the height, width and channel number of output from the last layer respectively. The *C'* denotes the number of channels for output.

Loss function: After performing the mixup data augmentation method, some label are not encoded in one-hot format. Therefore, to improve the learning ability of LungRN+NL on the distance between labeled vector and the predicted vector, and to achieve optimal trade-offs between sensitivity, S_e and specificity, S_p, we propose the following loss function:

$$\begin{aligned} loss = &- \frac{1}{N} \sum_{n=1}^{N} \sum_{c=1}^{4} [p_c y_{nc} \cdot log\sigma(x_{nc}) \\ &+ (1 - y_{nc}) \cdot log(1 - \sigma(x_{nc}))], \end{aligned} \tag{4}$$

where x_{nc} and y_{nc} refer to the c_{th} element in predicted vector and label vector of the n_{th} cycle, respectively. N is the total number of examples in each batch. p_c is the weight of solely positive examples to allow the LungRN+NL network to focus on adventitious lung sounds. It was used to trade of specificity and sensitivity.

4. Experiment Results

4.1. Experimental Setup and Evaluation Methods

Evaluation Methods: In this work, we have adopted the same evaluation method used in the official ICBHI 2017 contest [5]. The scoring method is defined as:

$$Sensitivity, S_e = \frac{P_c + P_w + P_b}{Crackle + Wheeze + Both}, \tag{5}$$

$$Specificity, S_p = \frac{P_n}{Normal}, \tag{6}$$

$$score = \frac{S_e + S_p}{2}, \tag{7}$$

where P_c, P_w, P_b and P_n are the number of correctly predicted records in four types of lung sounds, respectively. $Crackle$, $Wheeze$, $Both$ and $Normal$ are the total number of instances in each type of lung sound records, respectively.

Experimental Setup: In the experiments, we have implemented our LungRN+NL lung sound classification model with Pytorch in Python and tested it on our workstation with a 64-bit Linux machine with Intel i7-7800X 3.50GHz processor and an Nvidia GTX 2080TI graphics card.

We train the LungRN+NL model using an SGD optimizer with momentum and batch gradient descent [17]. The initial learning rate is 0.12 and it reduced to 0.1 times every 50 epoch. A 25% dropout rate is used to avoid over-fitting. A maximum batch size of 32 samples is used due to the limit in the GPU's memory.

4.2. Results and discussion

Evaluation of Data Augmentation Methods and Discussion: Table 2 illustrates the performance of our LungRN+NL with different data augmentation methods[1]. By comparing the performance of each data augmentation method, LungRN+NL has achieved the best sensitivity and score. The proposed mixup method has shown to be effective in identifying adventitious lung sound even without the use of traditional sound data augmentation methods.

Benchmark and Discussion: To make a comprehensive comparison with the state-of-the-art models [6–11], we have adopted two types of dataset division method: the official

[1] https://github.com/makcedward/nlpaug

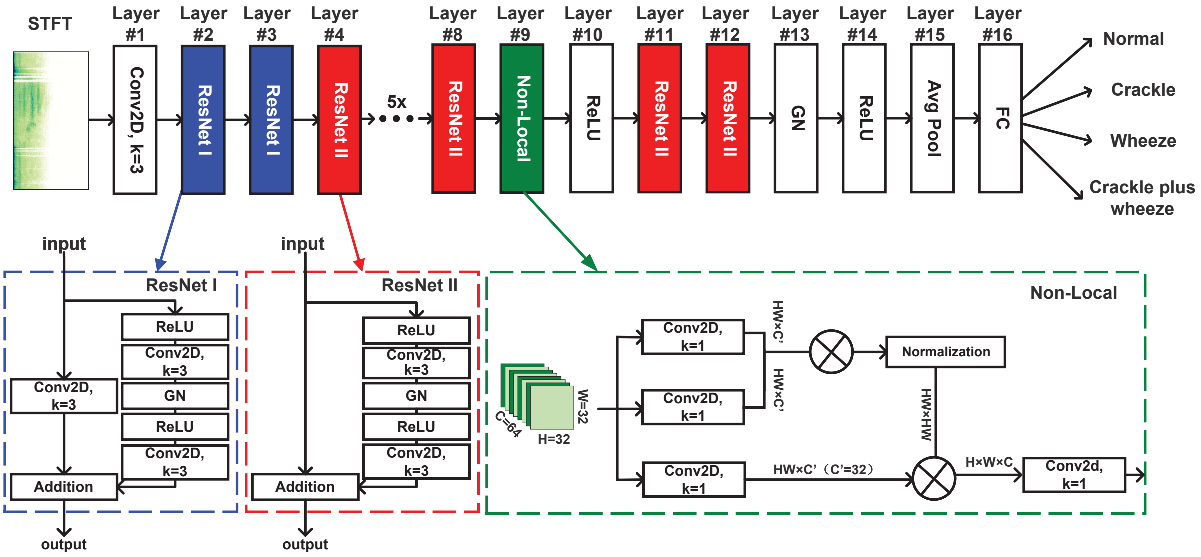

Figure 3: *Our proposed LungRN+NL neural network architecture. (a) An overview of the proposed architecture (b) The details of ResNet-I, ResNet-II and non-local layers used in LungRN+NL.*

Table 2: *Performance comparison for LungRN+NL using different data augmentation methods. The augmentation methods are: (a) crop audio's segment randomly; (b) Adjust audio's volume; (c) Inject noise; (d) Mask audio's segment; (e) Adjust audio's pitch; (f) Adjust audio's speed.*

Method	S_e	S_p	**Score**
Crop	38.09%	48.33%	43.21%
Loudness	30.98%	64.14%	47.56%
Noise	29.84%	71.16%	50.50%
Mask	33.33%	68.70%	51.02%
Pitch	31.99%	61.51%	46.75%
Speed	34.01%	67.93%	50.97%
LungRN + NL	**41.32%**	63.20%	**52.26%**

Table 3: *Comparison with state-of-the-art models*

	Reference	S_e	S_p	**Score**
Official[1]	[6]	-	-	39.56%
	[7]	20.81%	78.05%	49.43%
	[8]	-	-	49.86%
	[11]	31.12%	69.20%	50.16%
	LungRN+NL	**41.32%**	63.20%	**52.26%**
5 Fold[2]	[9]	58.43%	73.00%	65.70%
	[10]	48.63%	84.14%	66.38%
	LungRN+NL	**63.69%**	64.73%	64.21%

[1] Initial learning rate = 0.12, dropout = 0.25.
[2] Initial learning rate = 0.15, dropout = 0.15.

method [5] and the 5-fold cross-validation [9, 10]. The comparison results are presented in Table 3. Column S_e, S_p and *score* refer to the sensitivity, specificity and score reported in the state-of-the-art models [6–11] and our evaluation result from our LungRN+NL model. LungRN+NL achieves S_e, S_p and *score* of 41.32%, 63.20% and 52.26%, respectively. LungRN+NL has made an improvement of 2.1-12.7% in the official method compared to the state-of-the-art models, respectively [6–8, 11]. Since S_e and S_p refer to the network's ability to recognize adventitious lung sound and the network's ability to identify the normal sound, respectively, our method has improved S_e by 10.2% and 15.06% compared to the best state-of-the-art model [11] and [10], and achieved the best result in identifying adventitious lung sound.

5. Conclusions

In this work, we propose LungRN+NL, which is a ResNet neural network architecture a with non-local layer to perform classification on the adventitious lung sound ICBHI2017 dataset. To address the imbalanced problem, we have proposed to augment the adventitious lung sound classes using the mixup method. Based on the official ICBHI2017 scoring standards, we provide conclusive evidence that the LungRN+NL has achieved a performance score of 52.26%, which is improved by 2.1-12.7% compared to the state-of-the-art models [6–11]. However, we find that there are rooms for improvement to achieve higher accurate classification results.

6. Acknowledgements

This research is supported in part by the National Key Research and Development Program of China under Grant No. 2019YFB2204500 and in part by the Science, Technology and Innovation Action Plan of Shanghai Municipality, China under Grant No. 1914220370.

7. References

[1] V. Cukic, V. Lovre, D. Dragisic, and A. Ustamujic, "Asthma and chronic obstructive pulmonary disease (COPD)–differences and similarities," *Materia socio-medica*, vol. 24, no. 2, p. 100, 2012.

[2] A. Bohadana, G. Izbicki, and S. Kraman, "Fundamentals of lung

auscultation," *New England Journal of Medicine*, vol. 370, no. 8, pp. 744–751, 2014.

[3] J. Xie, Z. Tong, X. Guan, B. Du, H. Qiu, and A. S. Slutsky, "Critical care crisis and some recommendations during the COVID-19 epidemic in China," *Intensive care medicine*, pp. 1–4, 2020.

[4] W. H. Organization. Density of physicians. [Online]. Available: http://www.who.int/gho/health workforce/physicians density/en/

[5] B. M. Rocha, D. Filos, L. Mendes, G. Serbes, S. Ulukaya, Y. P. Kahya, N. Jakovljevic, T. L. Turukalo, I. M. Vogiatzis, E. Perantoni *et al.*, "An open access database for the evaluation of respiratory sound classification algorithms," *Physiological measurement*, vol. 40, no. 3, pp. 1–28, 2019.

[6] N. Jakovljević and T. Lončar-Turukalo, "Hidden Markov model based respiratory sound classification," in *Precision Medicine Powered by pHealth and Connected Health*, 2018, pp. 39–43.

[7] G. Chambres, P. Hanna, and M. Desainte, "Automatic Detection of Patient with Respiratory Diseases Using Lung Sound Analysis," in *International Conference on Content-Based Multimedia Indexing*, 2018, pp. 1–6.

[8] G. Serbes, S. Ulukaya, and Y. Kahya, "An automated lung sound preprocessing and classification system based onspectral analysis methods," in *Precision Medicine Powered by pHealth and Connected Health*, 2018, pp. 45–49.

[9] K. Kochetov, E. Putin, M. Balashov, A. Filchenkov, and A. Shalyto, "Noise Masking Recurrent Neural Network for Respiratory Sound Classification," in *International Conference on Artificial Neural Networks*, 2018, pp. 208–217.

[10] J. Acharya and A. Basu, "Deep Neural Network for Respiratory Sound Classification in Wearable Devices Enabled by Patient Specific Model Tuning," *IEEE Transactions on Biomedical Circuits and Systems*, pp. 1–1, 2020.

[11] Y. Ma, X. Xu, Q. Yu, Y. Zhang, Y. Li, J. Zhao, and G. Wang, "LungBRN: A Smart Digital Stethoscope for Detecting Respiratory Disease Using bi-ResNet Deep Learning Algorithm," in *IEEE Biomedical Circuits and Systems Conference (BioCAS)*, 2019, pp. 1–4.

[12] Y. Ma, X. Xu, Q. Yu, Y. Zhang, Y. Li, J. Zhao, and G. Wang, "Live Demo: LungSys - Automatic Digital Stethoscope System For Adventitious Respiratory Sound Detection," in *IEEE Biomedical Circuits and Systems Conference (BioCAS)*, 2019, pp. 1–1.

[13] H. Zhang and M. Cisse, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations*, 2018, pp. 1–13.

[14] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.

[15] X. Li, Y. Li, M. Li, S. Xu, Y. Dong, X. Sun, and S. Xiong, "A Convolutional Neural Network with Non-Local Module for Speech Enhancement," *Proc. Interspeech*, pp. 1796–1800, 2019.

[16] R. N. Bracewell and R. N. Bracewell, *The Fourier transform and its applications*. McGraw-Hill New York, 1986.

[17] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[18] R. Rollins, A. H. Marshall, E. McLoone, and S. Chamney, "Discrete Conditional Phase-Type Model Utilising a Multiclass Support Vector Machine for the Prediction of Retinopathy of Prematurity," in *IEEE 28th International Symposium on Computer-Based Medical Systems*, 2015, pp. 250–255.

[19] P. Luis and W. Jason, "The Effectiveness of Data Augmentation in Image Classification using Deep Learning," in *Computer Vision and Pattern Recognition*, 2017, pp. 1–8.

[20] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio Augmentation For Speech Recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015, pp. 1–4.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[22] L. O. Chua and T. Roska, "The CNN paradigm," *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, vol. 40, no. 3, pp. 147–156, 1993.