Mel-spectrogram and Deep CNN Based Representation Learning from Bio-Sonar Implementation on UAVs

2021 International Conference on Computer, Control and Robotics (ICCCR) | 978-1-7281-9035-8/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICCCR49711.2021.9349416

M. Hassan Tanveer
Robotics & Mechatronics Engineering
Kennesaw State University
Marietta, USA
mtanveer@kennesaw.edu

Hongxiao Zhu
Department of Statistics
Virginia Tech
Blacksburg, Virginia, USA
hongxiao@vt.edu

Waqar Ahmed
DITEN–University of Genova
PAVIS–Istituto Italiano di Tecnologia
Genova, Italy
waqar.ahmed@iit.it

Antony Thomas
DIBRIS
University of Genova
Genova, Italy
antony.thomas@dibris.unige.it

Basit Muhammad Imran
Mechanical Engineering
Virginia Tech
Blacksburg, USA
basit@vt.edu

Muhammad Salman
Department of Mechanical Engineering
Kennesaw State University
Marietta, USA
msalman1@kennesaw.edu

***Abstract*—In this paper, we present an approach for estimating the leaf density of trees while navigating in a forest. To this end, we consider an Unmanned Aerial Vehicle (UAV) equipped with a biosonar sensor that mimics the sonar sensors of echolocating bats. Such sensors provide a light-weight and cost-effective alternative to other widely used sensors such as camera, LiDAR and are gaining popularity among the robotics research community. The obtained echo signals during UAV navigation are processed to obtain the leaf density in the main lobe of the sonar first using a mel spectogram and then a Deep Convolutional Neural Network (CNN) trained on a set of known environment. We further evaluate our approach in simulation by considering trees with different leaf density (that is, resolution). It is seen that our method achieves promising results with an accuracy of 98.7%.**

***Index Terms*—Unmanned Aerial Vehicle, Deep Convolutional Neural Networks, Mel-Spectogram, Unknown Environment**

I. Introduction

Unmanned Aerial Vehicles (UAVs) have been used extensively in the recent past with applications spanning a variety of areas such as surveillance, search and rescue, mapping, and agriculture. The agility and three-dimensional mobility of these vehicles require sensors that provide information about the three-dimensional environment surrounding the vehicle [1]. Of late there has been an increased need for preserving rain forests [1] and its biodiversity. UAVs offer excellent capabilities in terms of mapping and exploring the biodiversity to harness the much needed information towards preserving the natural terrestrial ecosystems. In this paper, we develop an approach for navigation of UAVs among trees, shrubs and other complex foliage using biosonar sensors. In this regard, the echo signals are first converted to Mel-spectrogram and then fed to a Deep Convolutional Neural Network (DCNN) to classify the tree structure based on the leaf density (number of leaves in a given echo).

Bats are well known to fly and navigate through naturally complex and highly structured chaotic environments such as bushes and trees even in complete darkness [2]; similar is the behaviour of dolphins in the waters [3]. They achieve that by sending a chirp signal to the environment and perceiving the returned echo signal. Their ability to resolve two echo signals 2 ms apart and sense objects with a precision of 0.3 mm [4] enables them to not only *echolocate* themselves but also to classify objects according to their 3D structure, such as leaves of a tree. This motivates and serves as a strong foundation for the development of aerial robots embedded with bio-mimetic sonars or biosonars for aerial surveillance. The aerial navigation capability of UAVs is acquired by successful deployment of airborne sonars able to classify and explore complex features in chaotic environments.

Developing flying robots that mimic the sensing behavior of bat like species has been researched in the past [5], [6]. Yet, these works are tailored to developing mechanisms and structures which enable these robots to achieve the agility and flight performance of bats in a mechanical perspective. However, there is a need to address the sensing capabilities, especially light weight sensors that aid in mapping and exploration. This paper intends to address this gap. An overview of the proposed approach can be seen in Fig. 1.

As discussed previously, we classify the acquired echo signals to esimate the leaf density. State-of-the-art methods in audio signal classification use techniques such as dictionary

[1]https://www.xprize.org/articles/xprize-announces-new-10-million-competition

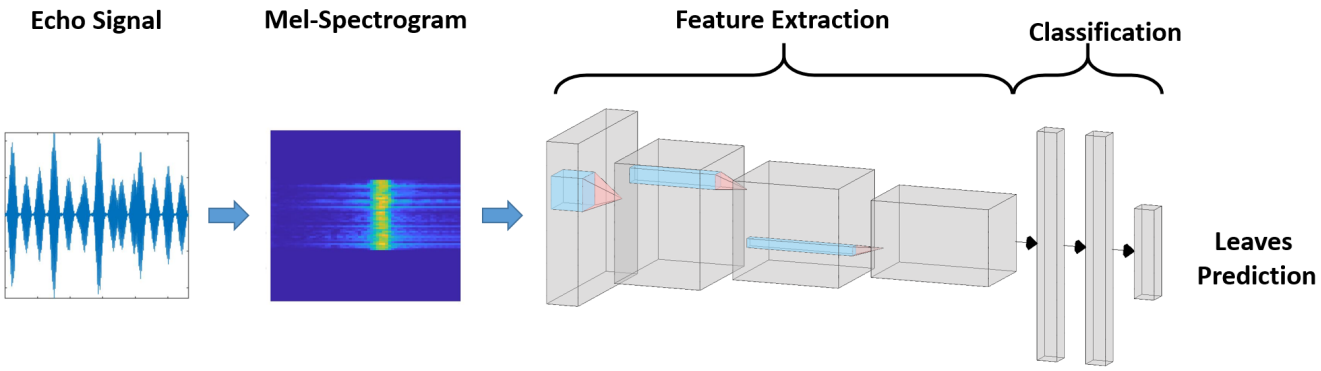


Fig. 1: The proposed pipeline. Mel-spectrogram converts the audio signal to an equivalent image. This image is then used for feature extraction using CNN. The number of leafs in the biosonar lobe is then returned as the output.

learning [7], wavelet-filterbank [8], [9], and most recently Convolutional Neural Networks (CNNs) [10], [11]. CNNs provide a powerful framework for the classification and feature extraction purposes. Even though CNNs have primarily been used primarily in the visual recognition context, CNNs have also found applications in diverse areas of engineering such as image processing [12], speech recognition [13]–[16], sports analytics [17], traffic signs classification [18], pedestrian detection [19], electron microscopy image processing [20], house number recognition [21] and robotics [22]. CNNs are very suited to our purpose due to 1) They capture energy modulation precisely when they are given inputs like Mel-spectrogram [23], 2) They are immune to noisy inputs unlike other approaches such as Mel-Frequency Cepstral Coefficients (MFCC) that are susceptible to noise [24].

Mel-spectrogram has been demonstrated to powerfully represent audio signals pictorially as inputs to CNNs [25]–[27]. This motivates us to use Mel-spectrogram as an input to CNN for our purpose, that is, *feature extraction from biosonar echos in aerial robots*. In particular, we use an approach similar to that presented in [28] for estimating the tree structure, that is, the leaf density. We first construct a Mel-spectrogram from the echo signals of the biosonar and then use a CNN to extract the required features of the environment. The rest of the paper is organized as follows. Section II details the complete methodology used to extract the number of leaves encountered by the sensor. In Section III, we discuss our experimental findings and the effectiveness of our method. Section IV concludes this paper.

II. Materials and Methods

We consider an aerial robot with the embedded biosonar that navigates in a forest with complex foliage. We note here that our approach is not restrictive to any particular environment and can be readily adapted to any given environment. As the UAV navigates, chirp signals are sent to the trees or to other vegetation which results in echos being returned.

Algorithm 1: Deep CNN training and Validation

Result: Leaf Predictions

A. Signal Pre-processing

if *Training* **then**
- 1) Apply Mel-Spectogram to Inputs
- 2) Apply DCNN to fit the model on training data (960 samples, 80% randomly selected)
- 3) Get the trained network

else
- Validate the DCNN of test set of 240 samples (20% randomly selected);

end

Make predictions and estimate model accuracy;

The echo returned consists of the reflections of chirp signal from various number of leaves whose density we intend to recover. An approach for simulating foliage echos that can produce simulated echoes by mimicking bat's biosonar has been introduced in our previous work [29]. An approach for simulating naturally looking trees, including their branches, sub-branches, and leaves can be found in [30] and we use the same approach for simulating a forest.

Fig. 1 provides an overview of our system that computes the number of leaves in the main lobe of the UAV biosonar. At first, the echo signal that has been sensed by the biosonar is used to generate a Mel-spectrogram. Mel-spectrogram then converts the audio signal to an equivalent image. This is depicted in the figure as a heatmap where the yellow region shows the information contained (data) and the blue region represents the empty data areas. This heat map is then sent to the CNN for feature extraction, which then takes this image of sound and then applies it to the first layer of convolution combined with Rectified Linear Unit (ReLU), followed by the layer of pooling to prepare the data for classification. The classification first flattens the data and further maxpools it,

thereby resulting in the prediction of the number of leaves in the sonar lobe.

The echo signal produced by biosonar is an outcome of the variation in air pressure being observed over time. A sampling rate of 44.1kHz is used to transform analog data (echo) into digital form. However, as shown in the right column of Fig 2, the echo signal contains several dead regions (amplitude close to zero) between spikes which are not useful for leaves density estimation. Therefore, we remove all the data points less than or equal to 3×10^{-4} from the echo signal to get a finer form of input sample for representation learning.

Subsequently, we map the trimmed echo signal from the time domain to the frequency domain using the fast Fourier Transform considering 64 overlapping windowed segments. We convert the color dimension (amplitude) to decibels and map the y-axis (frequency) onto the mel scale to form the Mel-spectrogram of echo signal (see Fig 1). The mel scale is a non-linear transformation of frequency-scale based on the perception of spikes, so that two pairs of frequencies separated by some Δ in the mel scale are perceived as being equidistant.

In this study, Mel-spectrogram represents a time-frequency representation of an echo signal: the power spectral density $P(f,t)$. It is sampled into a number of points around equally spaced times t and frequencies f (on a Mel frequency scale (1)). We do that by applying a bank of overlapping triangular filters with frame-duration of 20 ms and hop-duration of 10 ms that computes the energy of the spectrum in each band (total 64 bands). The mel frequency scale is defined as:

$$f_{mel} = 2595 \times log_{10}(1 + \frac{Hz}{700}) \tag{1}$$

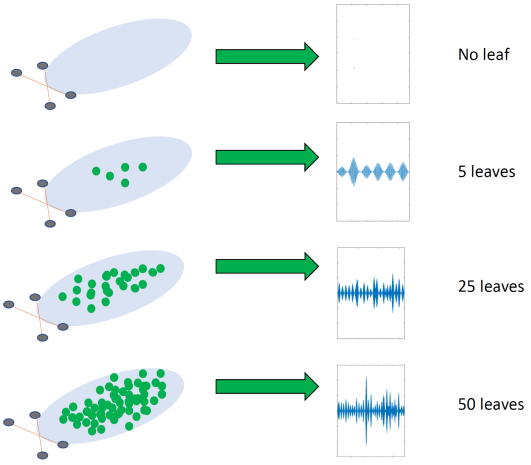


Fig. 2: The working mechanism of biosonar. The green dots represent the leaves inside the sonar main lobe. Ultrasonic pulses are emitted and the received echoes reflected from the leaves are shown on the right.

TABLE I: Deep CNN training Parameters used for all the 3 experiments i.e., with the resolution 5, 10, and 20.

CNN Parameters	Values
Training Set	70%
Validation Set	15%
Test Set	15%
Batch Size	128
Optimizer	adam
InitialLearnRate	1×10^{-3}
LearnRateDropFactor	0.1
LearnRateDropPeriod	10 epochs
MaxEpochs	30
Validation Frequency	20 Iterations
Shuffle	every-epoch
dropoutProb	0.2

B. Deep Convolutional Neural Network (CNN)

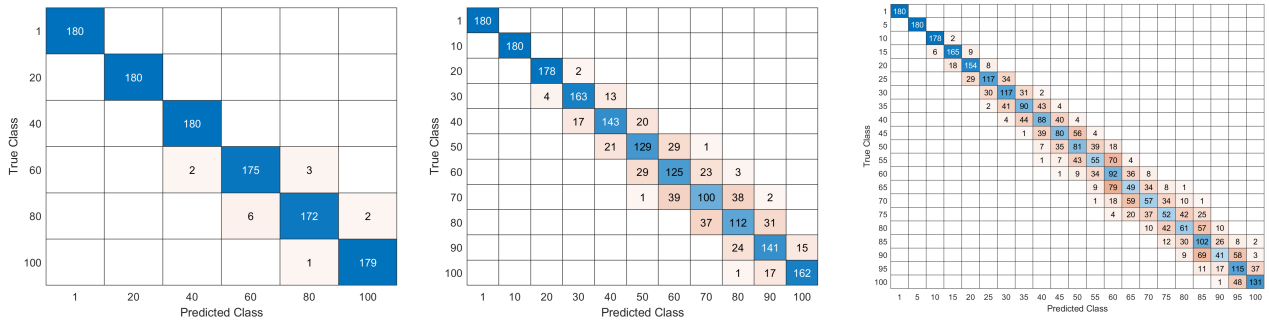
In deep learning, a deep Convolutional Neural Network (CNN) is used for representation learning of visual imagery. CNN is also known as shift invariant or space invariant artificial neural network, because of its shared-weights architecture and translation invariant characteristics. With the aid of deep convolutional neural networks, image understanding has achieved remarkable success in the past few years. Notable examples include residual networks [31] for image classification, FastRCNN [32] for object detection, and Deeplab [33] for semantic segmentation, to name a few.

In this work, we consider deep CNN, composed of a stack of convolutions, biases, fully-connected layers, and various pooling layers in which the output is a vector containing one score per class (e.g., number of leaves in the lobe). We number the parameterized layers of the Deep CNN $L = 1, ..., M + 1$. Each layer L corresponds to a convolution or fully-connected layer with input width I_L and output width O_L. In the case of a convolutional layer, I_L and O_L correspond to the number of input and output channels, respectively. We consider $L = M + 1$ to be the last layer of the network. Thus O_{M+1} is the size of the final output vector. The neural network is trained to minimize the cross-entropy loss:

$$\boldsymbol{\mathcal{L}}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} l\big(y^{(i)}, f(\mathbf{x}^{(i)}, \boldsymbol{\theta})\big) \tag{2}$$

where $\boldsymbol{\theta}$ are the parameters of model to be learned, $y^{(i)}$ is the ground truth label of example i, $f(\cdot)$ represents the activation function, $\mathbf{x}^{(i)} = \{x_1^{(i)}, x_2^{(i)}, \ldots, x_m^{(i)}\} \in \mathbb{R}^m$ denotes a training sample, and l is a loss measuring how well the neural network fits the data.

Specifically, we used a 5-layer CNN where each convolutional layer is followed by batch-normalization, ReLU as non-linearity function, and Max-pooling layer with stride of 2. To reduce the possibility of the network memorizing specific features of the training data, we added a small amount of dropout (0.2) to the input to the last fully connected layer. Finally, we employ softmax as an activation function to classify the numbers of leaves outputs.



(a) Resolution=20, Acc=98.70%

(b) Resolution=10, Acc=81.46%

(c) Resolution=5, Acc=57.80%

Fig. 3: DCNN confusion Matrix.

III. Results and Discussion

We consider a UAV equipped with a biosonar navigating in a simulated forest environment. To evaluate our proposed approach, we consider trees with varying leaf density. We vary the number of leaves that fall inside the main lobe of the sonar from 0 to 100 and evaluate three scenarios of different resolution, that is 5, 10, and 20. For example in scenario 1, the leaves are varied from 0 to 100 with an increment of 5 leaves. Each experiment was conducted using Signal Processing, Deep Learning, and DSP System Toolbox of MATLAB™ version R2019b. The overall dataset comprises of $25,200$ data samples representing 21 classes (1200 samples per class). As discussed in Sec II-A, the dead points from raw echo signal are removed to extract *trimmed* version of data samples. Consequently, echo samples of varied lengths (see Fig. 4), each representing a particular class, constitutes the proposed dataset. We kept the model training parameters fixed for all the three experiments. Each parameters and their specific details can be seen in Table I. The procedure for training and validation is elucidated in Algorithm 1.

However, the fixed data split ratio in each experiment corresponds to different number of classes and total data samples associated with it. The details are as follows,

Resolution=5. In this particular setting, we aim at classifying input sample as either 1, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95 or 100 leaves in the lobe. Consequently, we have 25,200 sample in total out of which randomly picked 17640, 3780 and 3780 samples are considered as training, validation and testing sets, respectively.

Resolution=10. In this setting, we aim at classifying input sample as either 1, 10, 20, 30, 40, 50, 60, 70, 80, 90 or 100 leaves in the lobe. Consequently, we have 13,200 sample in total out of which randomly picked 9240, 1980 and 1980 samples are considered as training, validation and testing sets, respectively.

Resolution=20. Similarly, in this setting, we aim at classifying input sample as either 1, 20, 40, 60, 80 or 100 leaves in the lobe. Consequently, we have 13,200 sample in total out of which randomly picked 5040, 1080 and 1080 samples are considered as training, validation and testing sets, respectively.

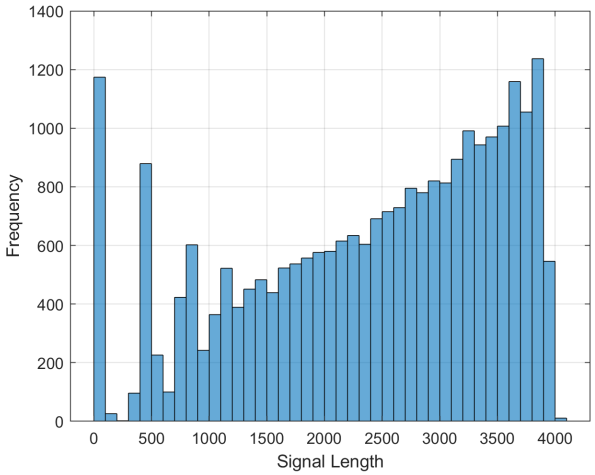


Fig. 4: Histogram of trimmed echo signals representing dataset used in experiments. There are total 25,200 samples representing 21 classes. The signal length is in milliseconds.

As shown in Fig. 3, our method showed most promising results with the accuracy of 98.7%. When the number of *true class* (leaves) are set through 1 to 40, and the resolution is set to 20 (increment from 1 to 40 is 20), we obtain 100% accuracy. However, as the *true class* is varied from 40 to 100, accuracy drops below 100. Average accuracy for *true class* 1 to 100 is 98.7%. Similarly, when the resolution is set to 10, the average accuracy when the *true class* is varies through 1 to 100 is 81.46%. However, if the resolution is decreased to 5, as shown by the confusion matrix the average accuracy drops to 57.8%.

IV. Conclusion

We have discussed an approach for classifying biosonar feedback of UAVs while navigating in forest and other complex foliage. The echoes are fed through a deep CNN fused with Mel-spectogram points to extract the leaf density. In this

way the number of leaves in the biosonar lobe is estimated. The approach is evaluated in simulation by varying the leaf density with different resolution. As can be confirmed from results, the greater the number of resolution with in the training results gives much better accuracy that is 97.80%. If resolution number is small, the accuracy turns out to be 57.80% or lower. Hence, for extraction of the features in the environment, we propose to use the resolution to a higher number.

REFERENCES

[1] V. Kumar and N. Michael, "Opportunities and challenges with autonomous micro aerial vehicles," *The International Journal of Robotics Research*, vol. 31, no. 11, pp. 1279–1291, 2012.

[2] J.-E. Grunwald, S. Schörnich, and L. Wiegrebe, "Classification of natural textures in echolocation," *Proceedings of the National Academy of Sciences*, vol. 101, no. 15, pp. 5670–5674, 2004. [Online]. Available: https://www.pnas.org/content/101/15/5670

[3] X. Qing, S. Liu, G. Qiao, Y. Dong, S. Ma, and D. He, "Acoustic propagation investigation of a dolphin echolocation pulse at water-sediment interface using finite element model," in *2018 OCEANS - MTS/IEEE Kobe Techno-Oceans (OTO)*, 2018, pp. 1–4.

[4] J. A. Simmons, N. Neretti, N. Intrator, R. A. Altes, M. J. Ferragamo, and M. I. Sanderson, "Delay accuracy in bat sonar is related to the reciprocal of normalized echo bandwidth, or q," *Proceedings of the National Academy of Sciences*, vol. 101, no. 10, pp. 3638–3643, 2004. [Online]. Available: https://www.pnas.org/content/101/10/3638

[5] A. Ramezani, X. Shi, S. Chung, and S. Hutchinson, "Bat bot (b2), a biologically inspired flying machine," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*, 2016, pp. 3219–3226.

[6] A. Ghanbari, E. Mottaghi, and E. Qaredaghi, "A new model of bio-inspired bat robot," in *2013 First RSI/ISM International Conference on Robotics and Mechatronics (ICRoM)*, 2013, pp. 403–406.

[7] J. Salamon and J. P. Bello, "Unsupervised feature learning for urban sound classification," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 171–175.

[8] ——, "Feature learning with deep scattering for urban sound analysis," in *2015 23rd European Signal Processing Conference (EUSIPCO)*, 2015, pp. 724–728.

[9] J. T. Geiger and K. Helwani, "Improving event detection for audio surveillance using gabor filterbank features," in *2015 23rd European Signal Processing Conference (EUSIPCO)*, 2015, pp. 714–718.

[10] E. Cakir, T. Heittola, H. Huttunen, and T. Virtanen, "Polyphonic sound event detection using multi label deep neural networks," in *2015 International Joint Conference on Neural Networks (IJCNN)*, 2015, pp. 1–7.

[11] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, 2015, pp. 1–6.

[12] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," *Neural Information Processing Systems*, vol. 25, 01 2012.

[13] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 6645–6649.

[14] L. Deng, O. Abdel-Hamid, and D. Yu, "A deep convolutional neural network using heterogeneous pooling for trading acoustic invariance with phonetic confusion," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 6669–6673.

[15] O. Abdel-Hamid, A. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1533–1545, 2014.

[16] T. N. Sainath, A. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for lvcsr," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 8614–8618.

[17] W. Ahmed, M. Amjad, K. Junejo, T. Mahmood, and A. Khan, "Is the performance of a cricket team really unpredictable? a case study on pakistan team using machine learning," *Indian Journal of Science and Technology*, vol. 13, no. 34, pp. 3586–3599, 2020.

[18] D. Cireşan, U. Meier, J. Masci, and J. Schmidhuber, "Multi-column deep neural network for traffic sign classification," *Neural Networks*, vol. 32, pp. 333 – 338, 2012, selected Papers from IJCNN 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0893608012000524

[19] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. LeCun, "Pedestrian detection with unsupervised multi-stage feature learning," 2012.

[20] D. Cirean, A. Giusti, L. M. Gambardella, and Schmidhuber, "Deep neural networks segment neuronal membranes in electron microscopy images," *Proceedings of Neural Information Processing Systems*, vol. 25, 01 2012.

[21] I. J. Goodfellow, Y. Bulatov, J. Ibarz, S. Arnoud, and V. Shet, "Multi-digit number recognition from street view imagery using deep convolutional neural networks," 2013.

[22] Z. Wang, Z. Li, B. Wang, and H. Liu, "Robot grasp detection using multimodal deep convolutional neural networks," *Advances in Mechanical Engineering*, vol. 8, no. 9, p. 1687814016668077, 2016.

[23] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, 2017.

[24] C. V. Cotton and D. P. W. Ellis, "Spectral vs. spectro-temporal features for acoustic event detection," in *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2011, pp. 69–72.

[25] D. Stowell and M. D. Plumbley, "Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning," *PeerJ*, vol. 2, p. e488, Jul 2014. [Online]. Available: http://dx.doi.org/10.7717/peerj.488

[26] S. Dieleman and B. Schrauwen, "Multiscale approaches to music audio feature learning," in *ISMIR*, 2013.

[27] J. Salamon and J. P. Bello, "Unsupervised feature learning for urban sound classification," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 171–175.

[28] G. de Magistris, P. Stinco, J. R. Bates, J. M. Topple, G. Canepa, G. Ferri, A. Tesei, and K. le Page, "Automatic object classification for low-frequency active sonar using convolutional neural networks," in *OCEANS 2019 MTS/IEEE SEATTLE*, 2019, pp. 1–6.

[29] M. Tanveer, A. Thomas, X. Wu, R. Mueller, P. Tokekar, and H. Zhu, "Recreating bat behavior on quad-rotor uavs—a simulation approach," in *The Thirty-Third International FLAIRS Conference (FLAIRS-33)*, 2020.

[30] M. H. Tanveer, A. Thomas, X. Wu, and H. Zhu, "Simulate forest trees by integrating l-system and 3d cad files," *arXiv preprint arXiv:2001.04530*, 2020.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[32] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[33] L.-C. 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 transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.