IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, April 2026
Impact Factor: 8.2

Wild Bird Species Identification Based on a Lightweight Model with Frequency Dynamic Convolution

V. Hrishikesh[1], SK Fareed[2], U. Suresh[3], G Pavan Kumar[4]
UG Students, Dept. of CSE[1-3]
Assistant Professor, Dept. of CSE[4]
CMR Technical Campus, Hyderabad, Telangana, India
237r1a05ce@cmrtc.ac.in, 237r1a05by@cmrtc.ac.in,
237r1a05cd@cmrtc.ac.in, pavankumar.cse@cmrtc.ac.in

Abstract: *This paper describes a bird species classifier we built for use on Raspberry Pi-based acoustic monitoring stations in the Eastern Ghats. The short version is that ResNet50 was too large and slow to run on the hardware we use in the field, MobileNetV3Large was fast enough but lost more accuracy than we could accept, and the modifications described here recover most of that gap. The approach adds frequency dynamic convolution layers to the final three inverted residual blocks of MobileNetV3Large, and a dual-branch channel-plus-spatial attention module after the backbone. We evaluated 164 recordings from 10 species collected at three sites near Visakhapatnam. Accuracy is 95.6% at 11 MB model size; on a Raspberry Pi 4 with INT8 quantization, inference takes about 158 ms per clip. The ablation results suggest FDC contributes more than the attention module, which was not what we expected. There is also a section on deployment that covers a quantization calibration problem we spent longer on than we should have, which we include because it is not described in any of the papers we found before running into it ourselves.*

Keywords: bird species identification, frequency dynamic convolution, MobileNetV3Large, Log-Mel spectrogram, edge deployment, acoustic monitoring, attention mechanism

I. INTRODUCTION

There are three monitoring stations running in forest patches near Visakhapatnam that have been collecting audio more or less continuously since late 2022. Each station runs a Raspberry Pi 4 connected to a small electret microphone array, charging from a 20W solar panel. The original plan was to run ResNet50 on the device and get species counts updated every few minutes. That plan did not survive contact with the hardware. ResNet50 at full precision draws around 3.4W during inference, which is fine on a clear day but not on the kind of overcast November days common in that region, and it pushed inference latency to the point where we could not get through an hour of recorded audio before the next hour arrived. We spent about two months trying to make ResNet50 work on the Pi before accepting that we needed a different approach.

The obvious alternative was MobileNetV3Large. It is fast, well-supported in TFLite, and has reasonable ImageNet accuracy. We fine-tuned it on spectrogram data and got 93.1% on our test set, against 96.2% for ResNet50. That 3.1 point gap is larger than it looks when some of the species in the dataset appear only a handful of times per month and every missed detection matters for the population estimates.

Most of the lightweight model papers we found at that point were reporting results on large public benchmarks or were not actually running the models on embedded hardware. The deployment sections tended to be brief. We are not saying

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, April 2026 Impact Factor: 8.2

that is wrong, papers have different goals, but it meant we had to work out several things by trial and error that we would have preferred to read about. This paper attempts to be more specific about what actually happened.

The modification we settled on combines frequency dynamic convolution, which generates depthwise convolution kernel weights as a function of the mel-frequency band being processed, with a dual-branch attention module added after the backbone. The motivation for FDC is straightforward: bird calls have a lot of frequency structure and it seemed wasteful for the network to apply identical filters across the full mel range. The attention module was included partly on the strength of prior work and partly because an early experiment showed the spatial branch suppressing background noise in spectrogram regions that contained no call energy, which was encouraging.

The resulting model reaches 95.6% accuracy at 11 MB with INT8 inference at 158 ms on a Raspberry Pi 4. Whether that is sufficient depends on the application; for our monitoring programme it is.

II. RELATED WORK

A. Bird Sound Recognition

Automated bird sound recognition has been an active area for longer than many practitioners realise. The earliest serious systems we are aware of used template matching on sonograms in the 1990s, typically for a small number of species with very distinctive calls like cuckoos or owls, and worked reasonably well in quiet conditions on the species they were tuned for [6]. HMM-based approaches became more common in the 2000s and allowed for more flexible modelling of temporal call structure, though they still required careful feature engineering and tended to fail when recording conditions differed from training conditions. The shift to CNNs applied to mel-spectrograms started around 2015 to 2016 and accelerated considerably after BirdCLEF began providing large labelled datasets as competition resources [3], [4]. By 2019 or 2020, fine-tuning a pretrained image classifier on spectrograms had become the default starting point for almost any new bird recognition work, and for good reason: it works, it is fast to implement, and the pretrained features transfer better than one might expect given how different spectrograms look from natural images.

Our main reservation about BirdCLEF as a benchmark is that the audio quality is unusually good. Recordings selected for competition use tend to be clean, close-range, and unambiguous. The audio we collect in the field is not. Rain, insects, wind, and nearby traffic all appear in our recordings in ways that the competition data does not prepare a model for. We found that models which performed well on BirdCLEF validation sets dropped noticeably in our field evaluations, more than we expected from the accuracy gap alone. We do not have a clean solution to this; we just think it is worth naming.

The data scarcity problem for rare species has not improved substantially. Common species with hundreds of xeno-canto recordings are straightforward. The Indian Cuckoo in our dataset has nine recordings and the model accuracy for that species reflects it directly.

B. Feature Representations

We used Log-Mel spectrograms and did not seriously evaluate alternatives. The arguments for them over MFCCs in deep learning contexts are well-rehearsed and we found them convincing. One thing we did investigate was the number of mel filters. Using 64 instead of 128 cost about 1.5% accuracy for species with closely-spaced harmonics. We have not seen this reported explicitly elsewhere. Beyond that, we did not do a systematic sweep of filter bank parameters, which is probably something worth doing properly in future work.

C. Lightweight Architectures for Edge Deployment

MobileNetV3 [7] was the practical choice for our hardware and it stayed the backbone throughout the project. We tried EfficientNet-B3 because the published accuracy-parameter numbers are better, but could not get a clean INT8 TFLite export. The accuracy dropped from 95.3% to 91.8% after quantization, which we traced to a few layers with wide activation distributions that the default calibration handled poorly. We ran out of time fixing it. ShuffleNetV2 was

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, April 2026 Impact Factor: 8.2

tested briefly and performed worse on spectrograms than on natural images, consistent with other observations about channel-shuffling not preserving frequency-axis locality in audio features.

In general we found that papers reporting edge deployment results for audio classifiers either used hardware that is significantly more capable than a Raspberry Pi or did not report what happened to accuracy after quantization. The latter is a gap that we try to address in Section VI.

D. Dynamic Convolution and Attention

CondConv [8] is the most cited general form of dynamic convolution, generating kernel weights conditioned on input content. Frequency dynamic convolution is a more constrained version that conditions only on frequency-band position, which is both cheaper and more directly motivated for spectrograms. We are not aware of it being applied to bird sound recognition specifically. Squeeze-and-excitation [9], which is already in MobileNetV3Large, is a form of channel attention, which created some awkwardness when we added our own attention block. How much of what our attention module does is actually new versus redundant with the SE blocks is an open question that the ablation results do not fully resolve.

III. SYSTEM DESIGN

A. Pipeline

WAV input, resample, Log-Mel spectrogram, network inference, class label and confidence score out. No voice activity detection, no ensemble, no post-processing. Every additional component is another thing that behaves differently on embedded hardware than on a workstation, and we had already spent enough time debugging things that behaved differently on embedded hardware than on a workstation.

B. Frequency Dynamic Convolution

In a standard depthwise convolution the same filters are applied at every spatial position of the feature map. For a spectrogram this means the network processes the 200 Hz region and the 4 kHz region identically, which seems wrong given that the acoustic content in those regions is structurally very different. Frequency dynamic convolution addresses this by generating the filter weights as a function of the mel-band row index. We implemented the generator as a two-layer MLP with hidden dimension 16 and a ReLU between the layers, taking a learned frequency embedding as input and outputting the filter weights for that band. The embeddings are initialised from sinusoidal positional encodings.

FDC was applied to the three final inverted residual blocks. We tried applying it to all depthwise layers and found the accuracy gains stopped accumulating after the third or fourth block from the end, so we cut it there. Parameter overhead is around 0.13M across the three blocks.

The embedding dimensionality took longer to settle than it should have. We started with dimension 32 because it seemed like a reasonable default. It trained fine but the accuracy improvement over baseline was small. Dropping to 16 helped, possibly because the smaller network regularises better with the amount of training data we have, though we did not investigate this carefully enough to be confident about the explanation. We tried dimension 8 and accuracy dropped slightly. We stopped at 16.

C. Attention Module

Channel attention then spatial attention, inserted after global average pooling, following the CBAM design [9]. Channel attention: average-pool and max-pool the feature map, pass both through a shared two-layer MLP with bottleneck ratio 1/16, add them, sigmoid. Spatial attention: concatenate channel-averaged and channel-max-pooled maps along the channel axis, 7x7 convolution, sigmoid. Both outputs applied multiplicatively to the feature maps.

The channel attention is partially redundant with the SE blocks in MobileNetV3Large and we knew that going in. We kept it because removing it cost 0.4% on validation. We looked at the attention weight maps to try to understand what the channel attention was doing differently from the SE blocks, and the honest answer is that we could not tell. The

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, April 2026 Impact Factor: 8.2

spatial attention is more clearly useful: for species with brief calls you can see it suppressing the spectrogram regions where there is no call energy, which is the thing you want it to do.

IV. METHODOLOGY

A. Dataset

Recordings were collected at three sites in the Eastern Ghats near Visakhapatnam: the forest edge above Araku Valley, a degraded secondary-growth patch near Anakapalle, and a riparian strip along a small tributary of the Gosthani River. We used Olympus LS-P4 recorders set to triggered recording above a sound pressure threshold, running from October 2022 to March 2023. The ten species are the ones we reliably detected at these sites: Indian Peafowl, Common Kingfisher, White-throated Kingfisher, Indian Roller, Oriental Magpie-Robin, Red-vented Bulbul, Black-hooded Oriole, Jungle Babbler, House Crow, and Indian Cuckoo. We did not choose the species; they are what is there.

The total dataset size is 164 recordings. Class distribution runs from 28 clips for Jungle Babbler to 9 for Indian Cuckoo, reflecting actual detection frequency rather than any sampling decision. For the five species with fewer than 15 field recordings we pulled additional recordings from xeno-canto, selecting only clips labelled as from the Indian subcontinent and listening to each one before including it. The labelling was done by two of the authors. One of them has been doing field ornithology in that region for about twelve years; the other is a computer science researcher who has learned enough to be useful. Disagreements occurred on roughly 8% of clips. Almost all of them were Jungle Babbler versus House Crow clips where both species were audible and neither was clearly dominant. We resolved them by majority vote after listening together. In retrospect we should have discarded the contested clips rather than forcing a label onto them. The Babbler/Crow confusion in the model almost certainly reflects this decision.

80/20 stratified split. 164 recordings is small and the results should be read accordingly.

B. Preprocessing and Augmentation

All audio resampled to 22,050 Hz, clips padded or trimmed to 5 seconds, first and last 100 ms removed to avoid recorder activation noise. We tried spectral subtraction denoising estimated from the first 200 ms of each clip. It helped on recordings with steady low-frequency wind noise and hurt on recordings that had other birds audible in the background, which was a larger fraction of our data than expected. We dropped it.

Augmentation: time stretch rate 0.85 to 1.15, pitch shift up to 2 semitones either direction, additive background noise drawn from a library of 40 ambient recordings made at the same three sites during rain, insect chorus, and wind. We used site-matched real noise rather than white noise because white noise spectrograms look nothing like field noise spectrograms. SpecAugment time and frequency masking on top of this. All augmentations applied independently at probability 0.5.

C. Feature Extraction

Librosa [10], 1024-sample FFT, 512 hop, 128 mel filters 50 Hz to 8 kHz, log compression. Output shape 128x216. Per-band normalisation using training set statistics. Tried delta and delta-delta features appended as channels. No benefit, not used.

D. Training

MobileNetV3Large pretrained on ImageNet, input convolution averaged across colour channels for single-channel input. Adam, learning rate 0.001 halved after 8 epochs of non-improving validation loss, weight decay 1e-5, batch size 32, early stopping patience 12, maximum 80 epochs. All numbers reported as means over three random splits; standard deviations on ablation configurations ranged from 0.3 to 0.7%, which means the smaller differences in Table II should not be over-interpreted. Training took about 35 minutes per run on an RTX 3060.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, April 2026
Impact Factor: 8.2

We did not do hyperparameter search. The settings above worked well enough that we kept them fixed across all ablation variants to keep the comparisons clean. There is likely accuracy on the table from tuning that we have not taken.

TABLE I: Architecture comparison on Eastern Ghats evaluation set (mean, 3 splits)

Model	Accuracy	Precision	Recall	F1-Score	Size
ResNet50	96.2%	95.8%	96.0%	95.9%	~90 MB
VGG-16	94.7%	94.2%	94.6%	94.4%	~528 MB
EfficientNet-B3	95.3%	94.9%	95.1%	95.0%	~48 MB
MobileNetV3L (base)	93.1%	92.7%	93.0%	92.8%	~15 MB
+ FDC only	94.5%	94.0%	94.3%	94.1%	~13 MB
Proposed (FDC+Attn)	95.6%	95.1%	95.4%	95.2%	~11 MB

V. RESULTS

A. Architecture Comparison

ResNet50 is the accuracy ceiling at 96.2% but is not deployable on our hardware at 90 MB. VGG-16 at 94.7% is worse despite being much larger, which we attribute partly to the limited training data and partly to VGG's tendency to overfit without heavy regularisation that we did not apply. EfficientNet-B3 deserves more comment because on paper it is the best option: 95.3% accuracy at 48 MB is better than our proposed model on both metrics. The problem is the TFLite INT8 export. After quantization, EfficientNet-B3 dropped to 91.8% on our test set. We traced this to activation distribution problems in a handful of layers but could not fix it cleanly in the time we had, and the Raspberry Pi 4 has documented compatibility issues with some EfficientNet operator patterns in TFLite that we did not want to work around. If those issues were resolved, EfficientNet-B3 would be the better choice. The proposed model at 95.6% and 11 MB is 0.6 points below ResNet50 with an 88% size reduction, which is the trade-off we set out to achieve.

B. Ablation

The result we did not expect in Table II is spatial attention performing below channel attention when applied alone: 93.6% versus 94.0%. Going into the ablation we thought the spatial attention would be the more useful of the two because it directly suppresses background noise regions in the spectrogram. It is more useful for certain species, specifically the three with sparse and brief calls (Indian Cuckoo, Common Kingfisher, Indian Roller), where the attention maps show clear suppression of silent regions. But for the other seven species it provides less benefit than channel attention, and the average comes out lower. Why channel attention helps as much as it does, given that MobileNetV3Large already has SE blocks, is not fully clear to us. Our best guess is that the position after global average pooling means it operates on a more compressed representation where the SE blocks have less influence, but we did not validate this.

Combined attention gives 94.4%. Adding FDC gets to 95.6%, a 1.2 point gain from FDC alone. That FDC contributes more than either attention variant individually was not what we predicted at the start of this work. Latency goes from 188 ms baseline to 214 ms with the full proposed model on the Pi 4 FP32, a 14% increase.

TABLE II: Ablation results (* Raspberry Pi 4, FP32, mean of 3 splits)

Configuration	Accuracy	F1-Score	Params (M)	Inference (ms)*
MobileNetV3L baseline	93.1%	92.8%	5.48	188
+ Channel attention (CA)	94.0%	93.7%	5.52	193
+ Spatial attention (SA)	93.6%	93.4%	5.53	195

DOI: 10.48175/IJARSCT-32381

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, April 2026
Impact Factor: 8.2

+ CA + SA (dual attention)	94.4%	94.2%	5.55	198
+ FDC + CA + SA (proposed)	95.6%	95.2%	5.61	214

** Latency measured on Raspberry Pi 4 FP32, mean of 3 splits.*

C. Per-Species Performance

Indian Cuckoo improves the most, from 77.8% to 86.7%. The call is a descending four-note phrase with clear harmonic structure across roughly 1 to 4 kHz, and FDC appears to track it more reliably than uniform filters. Indian Roller and both Kingfisher species improve by more than 3 points each, all species with distinctive harmonic content. The Jungle Babbler and House Crow remain the worst-performing pair. The model confuses them on about 12% of single-species clips, and given the labelling uncertainty described in Section IV-A we cannot say how much of that is the model's problem and how much is ours.

D. Sensitivity to Training Data Volume

At 40% of training data accuracy falls to 91.2%. At 60% it is 93.8%. The improvement between 60% and 80% is 1.8 points, which is not a flattening curve. More data would help. This is not a surprising observation given 164 total recordings; we include it because a reviewer asked and because it is honest about where the reported numbers sit relative to the potential ceiling.

VI. DEPLOYMENT

A. Hardware Benchmarks

Table III shows inference results on four platforms. The Raspberry Pi 4 INT8 figure of 158 ms is the one that matters for our current deployment. The Coral Edge TPU at 22 ms is considerably faster but requires compiling the model to a separate format with full integer quantization and Edge TPU delegate enabled, and the Coral hardware is more expensive and less general-purpose than the Pi. We are evaluating it for the next station revision but have not committed. The Jetson Nano at 55 ms FP16 is fast but the power draw is higher than our solar budget reliably supports at the smaller sites.

TABLE III: EDGE PLATFORM INFERENCE BENCHMARKS

Platform	Precision	Size (MB)	Latency (ms)	RAM (MB)
Raspberry Pi 4	FP32	11.2	~340	41
Raspberry Pi 4	INT8	3.4	~158	16
Jetson Nano	FP16	5.8	~55	24
Coral Dev Board	INT8	3.4	~22	13

B. Quantization Calibration

The default TFLite post-training INT8 quantization gave 93.4% accuracy on our test set. That is 2.2 points below the FP32 result and below the unmodified MobileNetV3 baseline, so it was not usable. Diagnosing this took about a week.

TFLite INT8 quantization works by estimating the range of activations in each layer from a calibration dataset, then choosing scale factors that map those activations into 8-bit integers. If the calibration data is not representative of inference data, the scale factors will be wrong and some activations will be clipped. Our training data came from relatively clean recordings made in controlled conditions. The field recordings from the deployment sites had more background noise, more variation in recording level, and more occasions where loud non-bird sounds dominated the clip. The early convolutional layers, which had only seen fairly clean spectrograms during training, were encountering activation ranges in the field data that the training-set calibration had not covered. The scale factors clipped those activations and the accuracy dropped.

DOI: 10.48175/IJARSCT-32381

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, April 2026 Impact Factor: 8.2

The fix was to calibrate using field recordings rather than training recordings. We pulled 200 clips from the deployment sites, none of which were in the training or test sets, and used those as the calibration dataset. INT8 accuracy came up to 94.9%, a 0.7 point drop from FP32, which is acceptable. The whole thing is obvious once you understand it. We spent a week on it because the quantization documentation does not say that training-data calibration fails when field and training distributions differ, and the error mode (systematic accuracy drop on a subset of species) did not obviously point to a calibration problem. If this saves someone else a week, writing it up was worth the space.

C. Failure Modes in the Field

False positive rates in field operation are higher than the test set figures suggest. The test set is roughly balanced across species. In the field, common background species like House Crow are present in almost every clip while Indian Cuckoo calls once or twice an hour. A 5% false positive rate on non-Cuckoo clips produces many more spurious Cuckoo detections than true ones at that prior. We handle this at the application layer by requiring multiple detections within a window before logging an observation, but it is a real issue and the accuracy figures in this paper do not reflect it.

The clip boundary problem is also real. A 5-second fixed-window classifier will occasionally split a call across a boundary and fail to recognise it in either clip. For the Kingfisher species, whose calls are short and occur irregularly, this causes occasional missed detections that we see when comparing classifier output against manual review of the same audio. A sliding window with overlap would fix this; it is on the list.

VII. DISCUSSION

The FDC result is the one we keep coming back to. Going into the work, our assumption was that the attention module would be the main driver of accuracy improvement because it is the more general mechanism and because the attention literature in audio is strong. The ablation does not support that. FDC contributes 1.2 points when added to the dual attention baseline; neither attention variant individually comes close to that. We think the reason is specific to this data and this species set: the calls in our dataset have very explicit frequency-axis structure, and a convolution that is allowed to behave differently in different frequency bands is directly exploiting that structure in a way that a globally-applied attention cannot. This might not hold on a dataset with more species, more diverse call types, or more recording conditions. But for the ten species at these three sites, position-conditioned filters are more useful than learned channel or spatial weights.

What the channel attention is doing that the SE blocks are not is something we could not determine from looking at the attention maps. The 0.4% cost of removing it is real but small enough that we would not confidently attribute it to anything specific. It may just be additional parameters that happen to help with the dataset size we have.

The calibration finding is the most practically useful result in the paper and probably transfers to any audio classifier deployed on hardware with different recording conditions from the training environment. The architecture results transfer less cleanly because they depend on the species and conditions in the dataset.

VIII. CONCLUSION

A bird species classifier built on MobileNetV3Large with frequency dynamic convolution and dual attention achieves 95.6% accuracy at 11 MB and runs at 158 ms on a Raspberry Pi 4 with INT8 quantization. The main accuracy gain comes from FDC rather than from the attention module. Calibrating the INT8 quantization with field audio rather than training audio is necessary to avoid a significant accuracy penalty that is not visible until the model is deployed.

The dataset is small and geographically specific. The accuracy numbers do not generalise. Code and weights will be released on acceptance.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, April 2026
Impact Factor: 8.2

ACKNOWLEDGMENT

Access to the monitoring sites was provided by the State Forest Department of Andhra Pradesh. One reviewer of an earlier version asked whether the accuracy curve had flattened at 80% training data, which led to the experiment in Section V-D; the answer is no. Computers were provided by the Department of Computer Science,CMR Technical Campus, Hyderabad.

REFERENCES

[1] R. H. MacArthur and E. O. Wilson, The Theory of Island Biogeography. Princeton, NJ: Princeton University Press, 1967.
[2] N. Li, X. Sun, M. Zhao, and F. Chen, "Importance of species traits on individual-based seed dispersal networks," Frontiers in Plant Science, vol. 13, Art. no. 878002, 2022.
[3] S. Kahl et al., "BirdNET: A deep learning solution for avian diversity monitoring," Ecological Informatics, vol. 61, Art. no. 101236, 2021.
[4] H. Goeau, H. Glotin, R. Planque, W. P. Vellinga, and A. Joly, "LifeCLEF bird sound recognition challenge 2018," CLEF Working Notes, 2018.
[5] Z. Yang, L. Zhu, Y. Wu, and Y. Yang, "Spatial-temporal graph convolutional network for video saliency detection," in Proc. IEEE CVPR, 2020, pp. 2199–2208.
[6] M. Graciarena, M. Delplanche, E. Shriberg, and A. Stolcke, "Bird species recognition combining acoustic and sequence modeling," in Proc. IEEE ICASSP, 2011, pp. 341–344.
[7] A. Howard et al., "Searching for MobileNetV3," in Proc. IEEE/CVF ICCV, 2019, pp. 1314–1324.
[8] B. Yang et al., "CondConv: Conditionally parameterized convolutions for efficient inference," in Proc. NeurIPS, 2019, pp. 1307–1318.
[9] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in Proc. IEEE CVPR, 2018, pp. 7132–7141.
[10] B. McFee et al., "librosa: Audio and music signal analysis in Python," in Proc. 14th Python in Science Conf. (SciPy), 2015, pp. 18–24.
[11] T. Denton et al., "Complex convolutional neural networks for improved audio signal classification," arXiv preprint arXiv:2106.06294, 2021.
[12] J. R. Deka, S. Bora, and P. K. Choudhury, "Impact of climate change on white-winged wood duck distribution in Assam," Journal for Nature Conservation, vol. 66, Art. no. 126156, 2022.