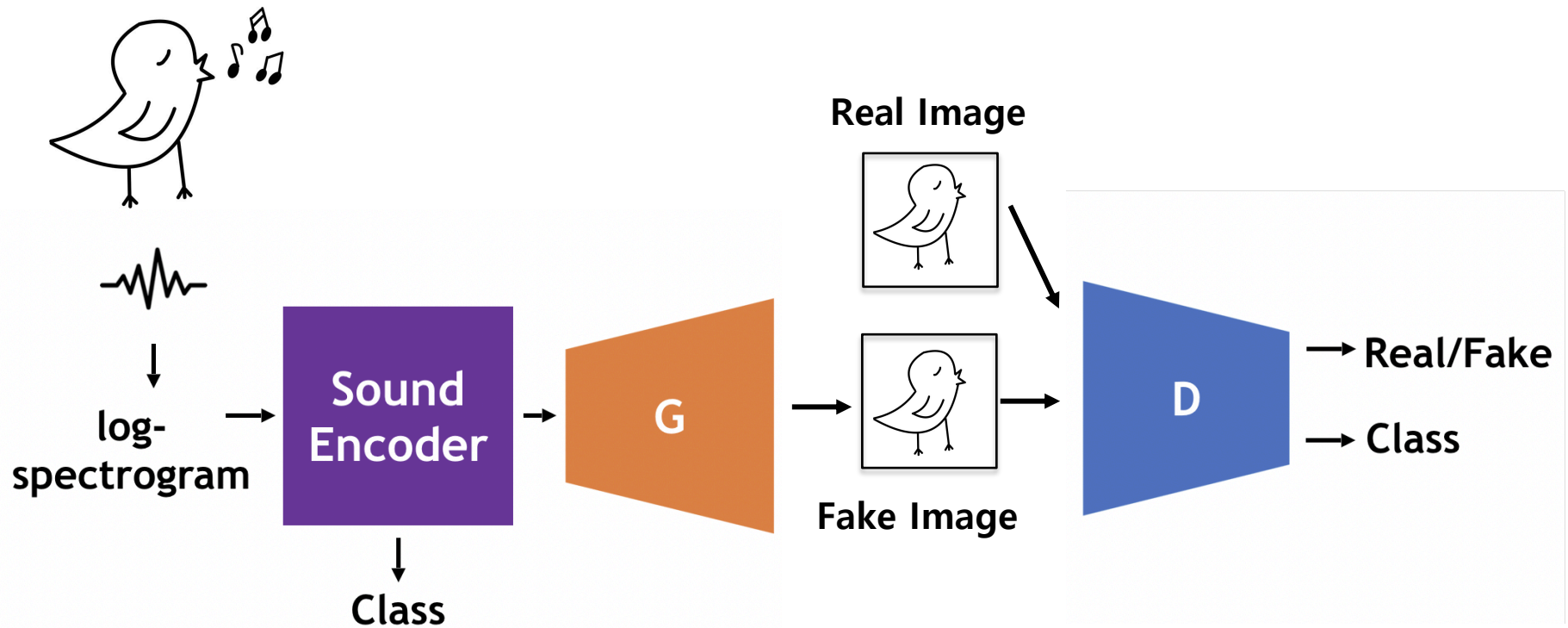# S2I-Bird: Sound-to-Image Generation of Bird Species using Generative Adversarial Networks.

Joo Yong Shim, Joogheon Kim and Jong-Kook Kim
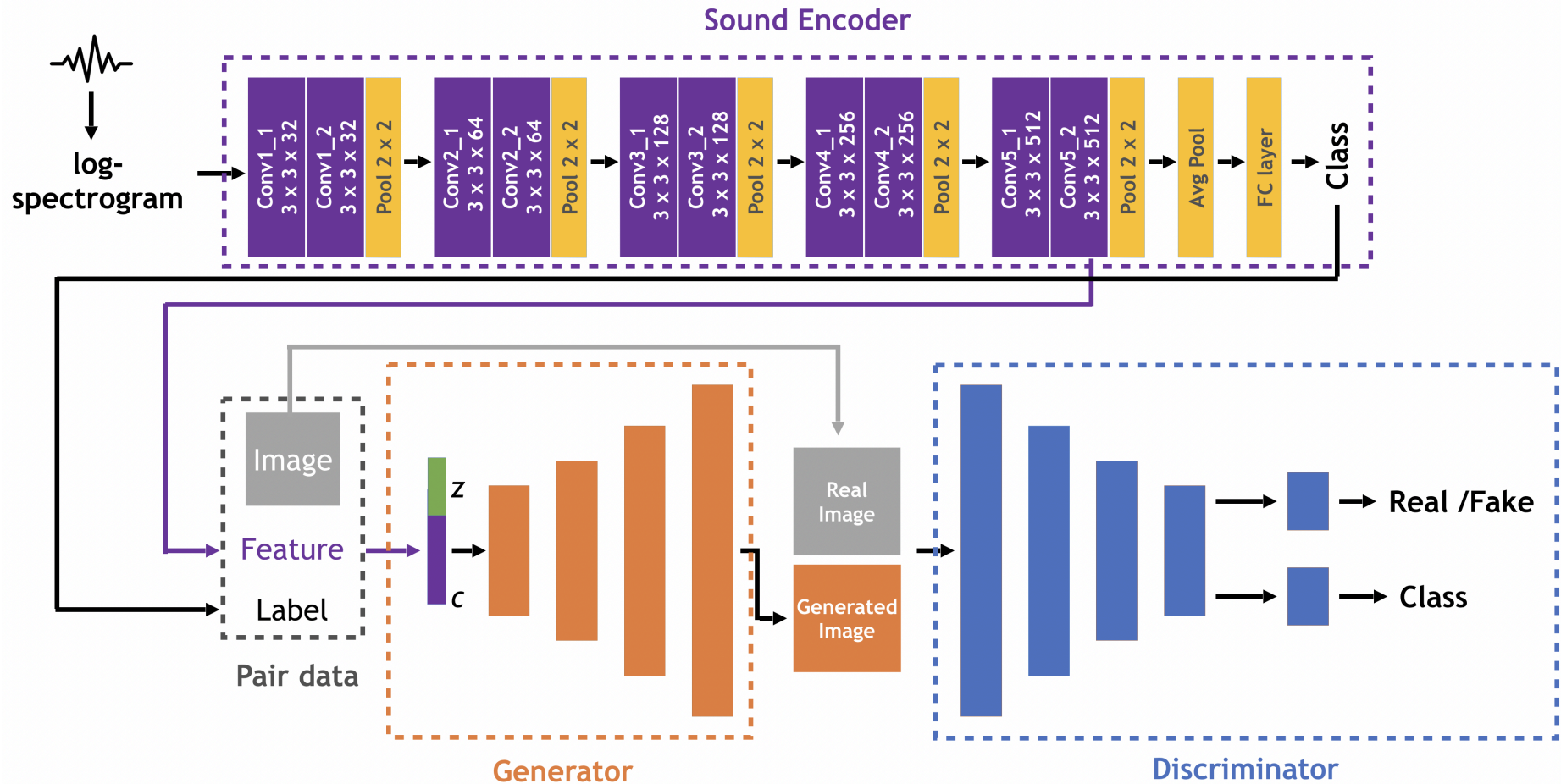
School of Electrical Engineering, Korea University
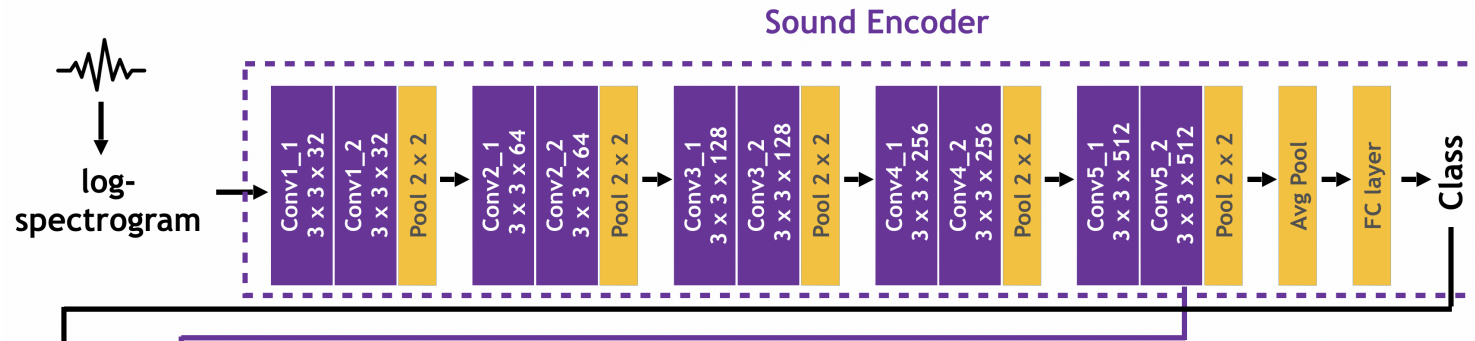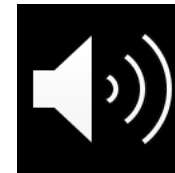
ICPR 2020

**High Performance Computing
& Computational Intelligence LAB**

Abstract

Model Architecture

Sound Encoder

| Conv1_1 3 x 3 x 32 | Conv1_2 3 x 3 x 32 | Pool 2 x 2 | Conv2_1 3 x 3 x 64 | Conv2_2 3 x 3 x 64 | Pool 2 x 2 | Conv3_1 3 x 3 x 128 | Conv3_2 3 x 3 x 128 | Pool 2 x 2 | Conv4_1 3 x 3 x 256 | Conv4_2 3 x 3 x 256 | Pool 2 x 2 | Conv5_1 3 x 3 x 512 | Conv5_2 3 x 3 x 512 | Pool 2 x 2 | Avg Pool | FC layer | Class |

log-spectrogram

Pair data: Image, Feature, Label

Generator: z, c, Real Image, Generated Image

Discriminator: Real /Fake, Class

# Model Architecture

Sound Encoder

Model Architecture
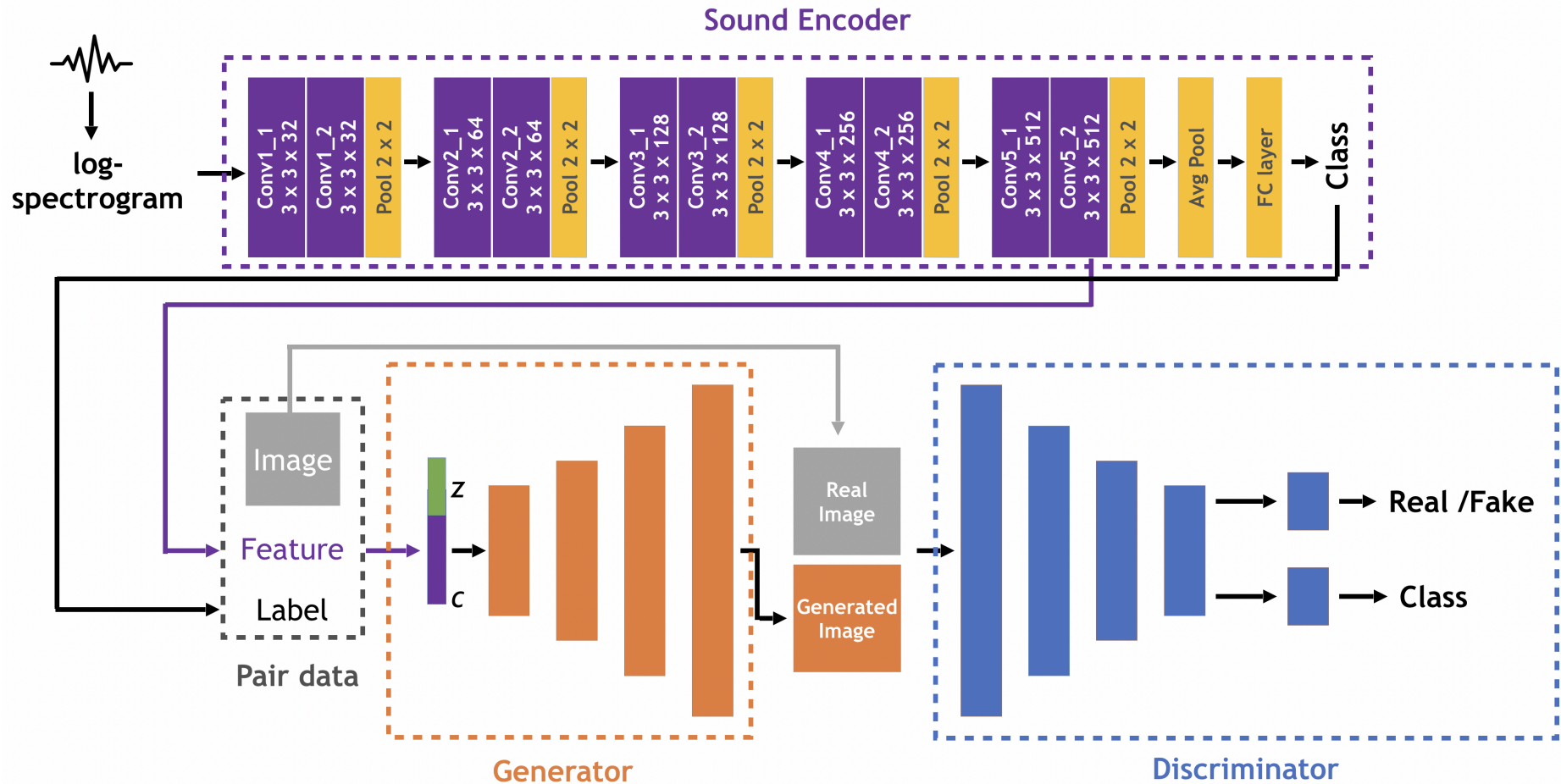
Dataset

Caltech- UCSD Birds-200-2011 (CUB) dataset
Xeno-Canto collaborated database

Species	Image	Log-spectrogram of Audio Features
Flycatcher	1, 254	4, 677
Gull	1, 407	2, 244
Kingfisher	1, 200	1, 223
Sparrow	1, 244	33, 358
Tern	1, 256	3, 800
Vireo	1, 227	9, 224
Warbler	1, 490	18, 293
Woodpecker	1, 397	2, 959
Wren	1, 257	13, 686

Performance Evaluation

Sound Encoder Evaluation
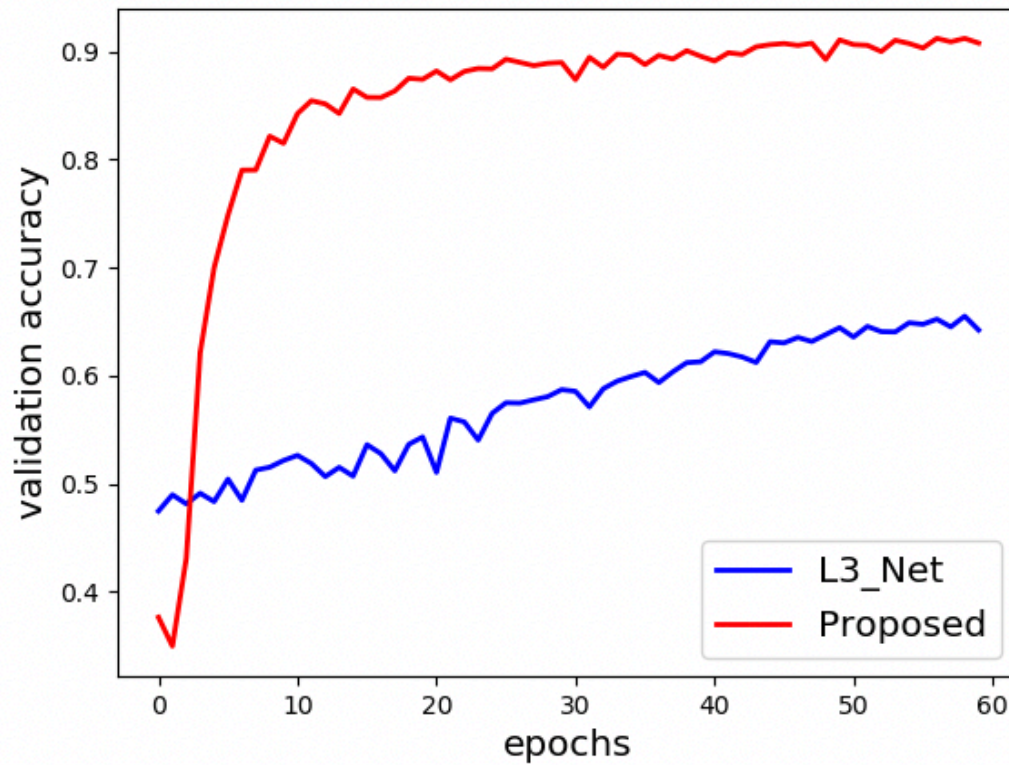
Performance Evaluation

Inception Score

Input condition	Inception Score
Upper Bound	4.10 ± 0.19
Label	2.81 ± 0.19
Mel-Spectrogram	1.76 ± 0.07
Our Sound Encoder	3.86 ± 0.31

Human Evaluation on generated images
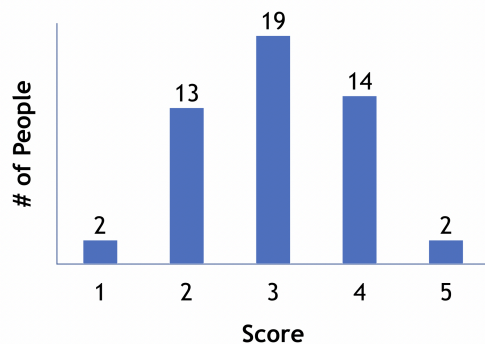
Label
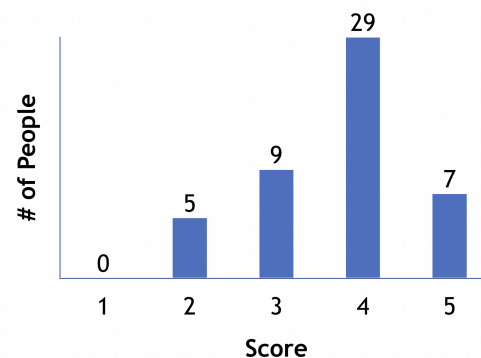
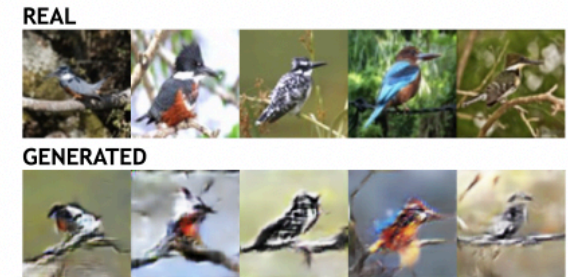Features from ours

Performance Evaluation

Generated images (64 × 64) on sound encoded feature condition.

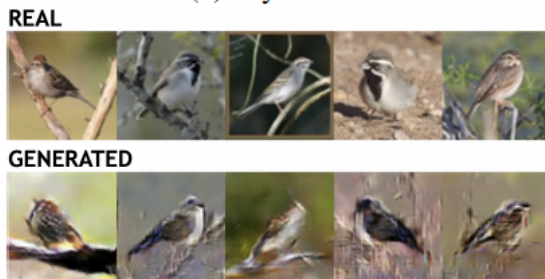(a) Flycatcher
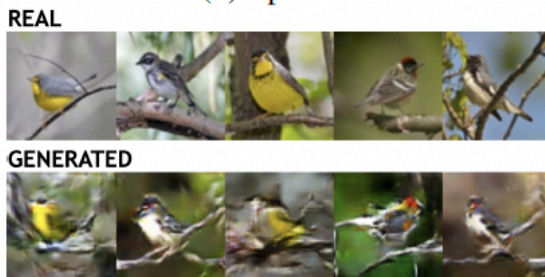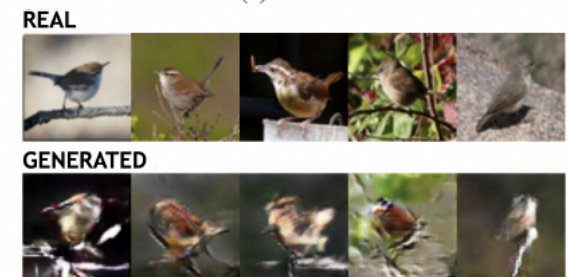
(b) Gull

(c) Kingfisher

(d) Sparrow

(e) Tern

(f) Vireo

(g) Warbler

(h) Woodpecker

(i) Wren

Thank you