The 38th Annual AAAI Conference on Artificial Intelligence
FEBRUARY 20-27, 2024 | VANCOUVER, CANADA

SoundCount: Sound Counting from Raw Audio with Dyadic Decomposition Neural Network

DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF OXFORD

Yuhang He[1], Zhuangzhuang Dai[2], Long Chen[3,4], Niki Trigoni[1], Andrew Markham[1]
[1]Department of Computer Science, University of Oxford, UK
[2]Aston University, UK. [3]Institue of Automation, Chinese Academy of Sciences, China. [4]WAYTOUS Ltd., China

1. Problem Definition

Given one-channel sound raw waveform, we aim to

1. count the sound event number.
2. regardless of sound class label, start/end time.

where,

1. acoustic scene is highly polyphonic.
2. inter/intra sound overlap in time/freq. domain.

Example: how many seagulls are heard in the audio?

2. Difference from SED

Sound Event Detection (SED) further

1. localize sound event's temporal position.
2. classify each sound event's semantic label.

Sound Count, instead,

1. count the present sound number (how many?).
2. Analogous to crowd counting in vision

3. Challenges in Sound Count

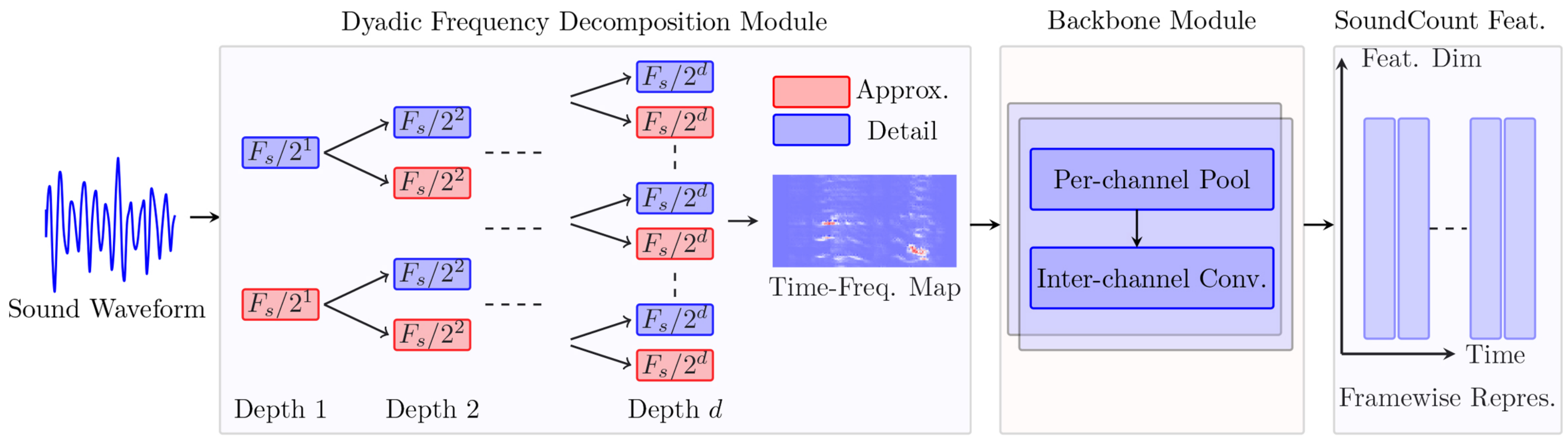
Learn a time-frequency (TF) map that can handle:

Challenge 1: *Loudness Variability*.

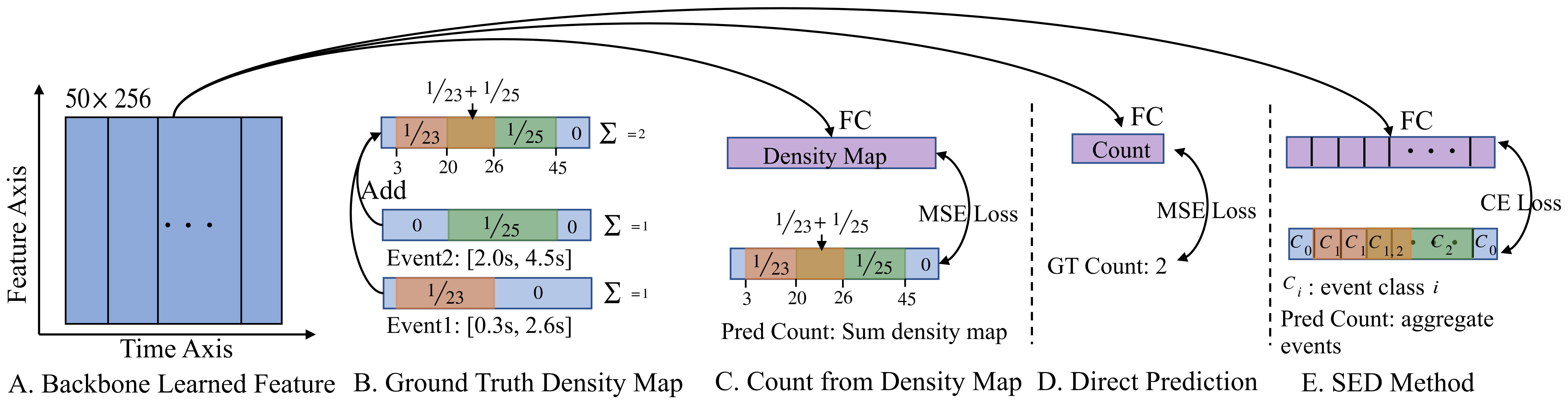
Challenge 2: *Frequency Overlap*.

Challenge 3: *Polyphonicity.*

4. Dyadic Decomposition Network

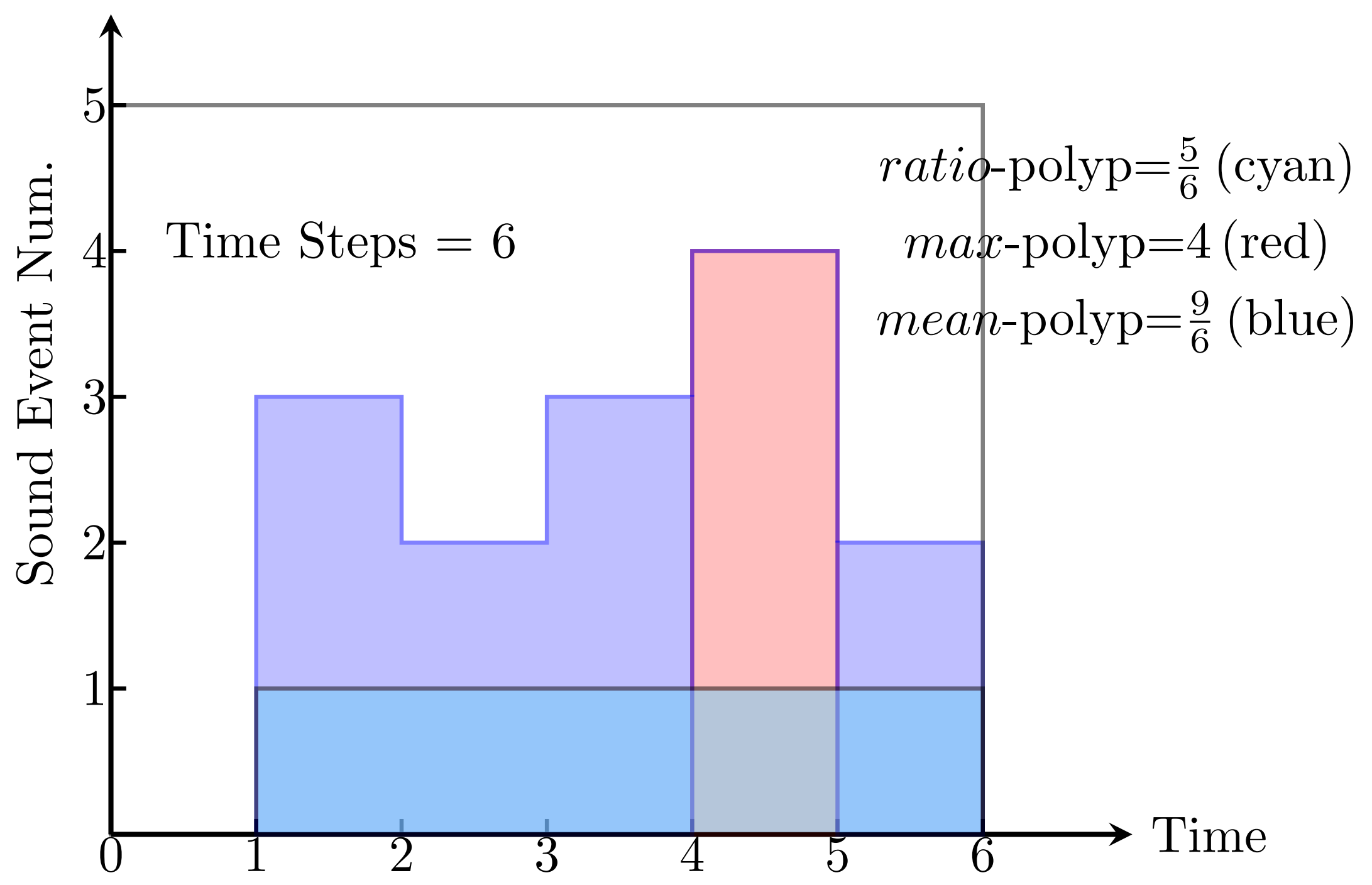


5. Density Map based Count



A. Backbone Learned Feature B. Ground Truth Density Map C. Count from Density Map D. Direct Prediction E. SED Method

6. Count Difficulty Quantifcation



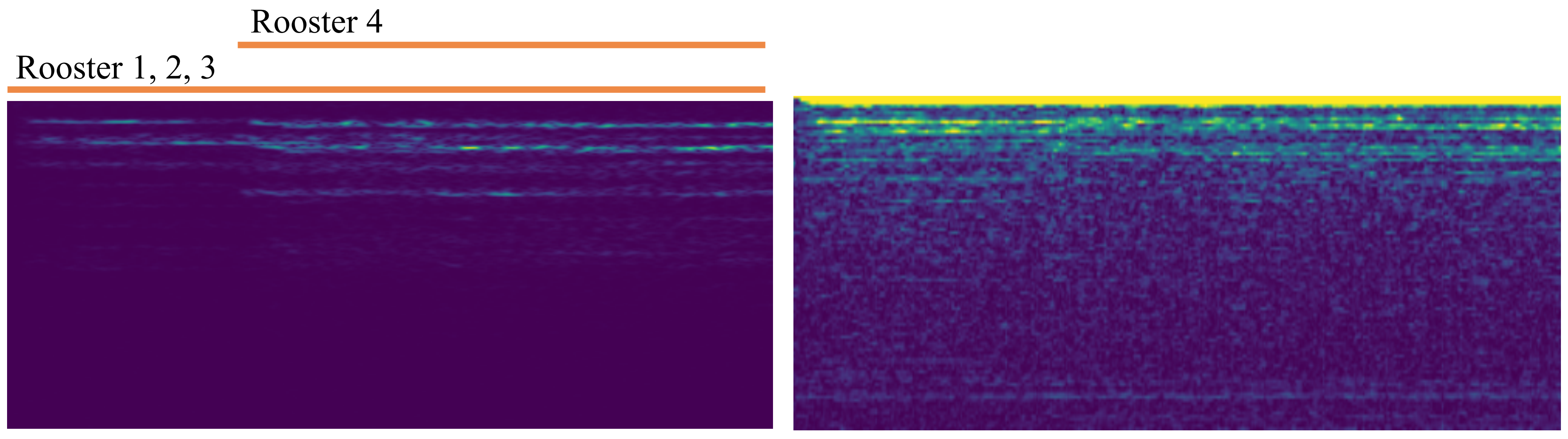
1. Polyphony Ratio
2. Max-Polyphonicity
3. Mean-Polyphonicity

7. Experiment Result

Dataset: five main categories: Bioacoustics, Indoor, Outdoor, Audio, Music.

Comparing Methods: two signal processing methods, three SED based methods, one source separation method

Experiment Result: DyDecNet is best-performing.



DyDecNet learned TF Feature Traditional MFCC Feature

Conclusion: 1. Split sound count from SED problem. 2. Propose a new TF map learning framework to handle the count challenge.