A Framework for Bird Sound Spectrogram Segmentation: Leveraging Deep Learning with Class Imbalance Mitigation

Shraddha Belbase Yeshiva University
sbelbase@mail.yu.edu

Abstract

Segmenting bird vocalizations from spectrograms presents unique challenges, as meaningful signals are often sparse and surrounded by environmental noise. This work introduces a ResNet-inspired U-Net framework, leveraging residual connections and the Focal Tversky Loss to address these challenges effectively. By extracting hierarchical features and mitigating class imbalance, the model enhances its ability to segment bird vocalizations amidst noisy backgrounds. Rigorous evaluations validate the robustness of our approach, achieving a mean IoU of 63.24% on a challenging bird sound segmentation dataset. This framework provides a foundation for advancing segmentation tasks in ecological monitoring and bioacoustic analysis.

1. Introduction

Image segmentation is a cornerstone task in computer vision, underpinning advancements in diverse fields such as medical diagnostics [6], environmental monitoring [9], and bioacoustic analysis [8]. This process involves partitioning an image into meaningful segments, typically corresponding to objects or regions of interest. Despite its importance, segmentation tasks often face challenges that compromise model performance and generalization, particularly when dealing with highly imbalanced datasets.

Class imbalance is a common issue in segmentation tasks where the foreground object of interest occupies a significantly smaller portion of the image compared to the background [5]. For instance, in medical imaging, tumor regions are often tiny compared to the surrounding healthy tissue [6]. Similarly, in ecological applications such as bird sound spectrogram segmentation, the vocalizations occupy a sparse and irregular distribution within the spectrogram [9]. This imbalance can bias models toward overfitting the background, thereby reducing the ability to identify critical foreground regions [4, 1].

Previous works have proposed various strategies to address this issue. Architectural innovations, such as U-Net [6], have been widely adopted due to their ability to retain spatial details through skip connections while capturing high-level features via an encoder-decoder structure. Loss functions such as the Focal Loss [4] and Tversky Loss [7] have been designed to handle class imbalance by penalizing false positives and false negatives differentially. These advancements have significantly improved segmentation performance, but challenges remain in generalizing to noisy and complex environments.

In this work, we build upon these advancements by introducing:

- A **ResNet-inspired U-Net architecture**, combining the strengths of residual connections and hierarchical feature extraction to enhance gradient flow and capture detailed spatial information.
- The **Focal Tversky Loss**, which emphasizes hard-to-segment regions and mitigates the effects of class imbalance.
- A robust evaluation framework to validate the proposed methodology on a bird sound spectrogram segmentation dataset.

The significance of this research lies in its application to bioacoustic analysis, a field where accurate segmentation of bird vocalizations from spectrograms is crucial for monitoring species diversity, understanding behavioral patterns, and supporting conservation efforts. By addressing the dual challenges of class imbalance and complex data distributions, this work aims to provide a reliable and generalizable framework for imbalanced segmentation tasks.

2. Related Work

Class imbalance in image segmentation has been addressed through various innovations, including architectural modifications, tailored loss functions, and data augmentation strategies.

2.1. Architectural Advances

U-Net, introduced by Ronneberger et al., is one of the most influential architectures for biomedical image segmentation [6]. Its encoder-decoder structure, combined with skip connections, preserves spatial information during upsampling. More recent advancements include DeepLabV3+ [2], which employs atrous convolutions to capture multi-scale context, and Vision Transformers [3], which leverage self-attention mechanisms for long-range dependencies. In bioacoustics, CNN-based encoder-decoder architectures have been used successfully for sound spectrogram segmentation [8].

2.2. Loss Functions for Imbalanced Data

Loss functions play a crucial role in handling class imbalance. Focal Loss [4] focuses on hard-to-classify examples, while Tversky Loss [7] balances false positives and false negatives. Focal Tversky Loss [1] combines these concepts to emphasize hard examples and is particularly effective in imbalanced segmentation tasks. Studies have shown its efficacy in medical imaging and environmental monitoring [7, 1].

2.3. Segmentation in Bioacoustics

Bird sound segmentation has gained increasing attention due to its potential in ecological research and wildlife conservation. Zhang et al. demonstrated the application of CNNs for isolating bird sounds from spectrogram data [9]. Recent approaches, such as the use of Vision Transformers [3], have shown promise in capturing complex patterns in spectrograms. However, challenges remain in handling variability in recording conditions and background noise [8].

3. Methods

3.1. Dataset Description

The dataset used in this study consists of spectrogram images generated from bird sound recordings and their corresponding binary masks. Each spectrogram represents a time-frequency analysis of audio signals, while the masks highlight regions containing bird vocalizations. These paired spectrogram-mask samples form the basis of the supervised segmentation task.

Dataset Structure The dataset is divided into the following subsets:

- **Training Set:** 1,000 spectrogram-mask pairs.
- **Validation Set:** 200 spectrogram-mask pairs.
- **Test Set:** 300 spectrogram-mask pairs.

Data Preprocessing To ensure compatibility with the model, the following preprocessing steps were applied:

1. **Grayscale Conversion:** Both spectrogram images and masks were converted to grayscale to simplify the data and focus on intensity values, which correspond to audio features.

2. **Resizing:** Each spectrogram and mask was resized to 128 × 128 pixels. Bilinear interpolation was used for spectrograms to preserve frequency-related details, while nearest-neighbor interpolation was applied to binary masks to maintain sharp edges.

3. **Normalization:** Pixel intensity values in spectrograms were normalized to the range [0, 1] for stable model training and optimization.

Foreground Composition Analysis A key characteristic of this dataset is the significant class imbalance between foreground (bird vocalizations) and background (ambient noise). As shown in Figure 1, the majority of spectrograms contain less than 10% foreground coverage. This imbalance underscores the challenges of segmenting sparse and irregularly distributed vocalization regions.

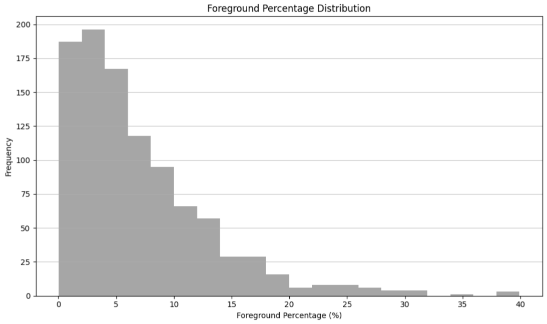

Figure 1. Distribution of foreground percentages in spectrogram masks.

Dataset Loading and Batching A custom PyTorch `Dataset` class was used to load and preprocess the data. The spectrograms and masks were read from their respective directories, paired lexicographically to ensure alignment, and processed as described above. The data pipeline employs the PyTorch `DataLoader` to create batches of size 16 for training, validation, and testing. Shuffling was applied to the training data to enhance model generalization.

Data Augmentation Although this study does not include augmentation techniques such as random rotations or brightness adjustments, future work could incorporate these strategies to improve generalization further. The focus here is on addressing challenges posed by the dataset's intrinsic class imbalance.

Dataset Visualization Representative examples of preprocessed spectrogram images and their corresponding masks are shown in Figure 2, highlighting the diversity and complexity of the dataset.

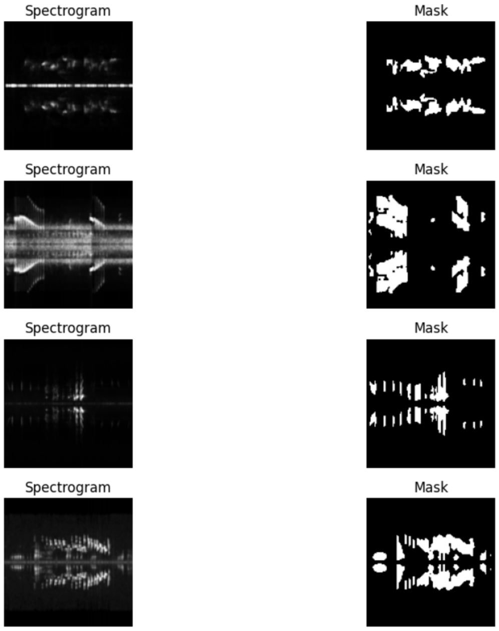

Figure 2. Examples of preprocessed spectrograms and their corresponding binary masks.

3.2. Model Architecture

The architecture combines a ResNet-inspired encoder with a U-Net-style decoder. The encoder uses residual connections to facilitate gradient flow and capture hierarchical features. Each encoder block consists of two convolutional layers followed by batch normalization and ReLU activation. Dropout layers are included in the bottleneck with a rate of 0.3 to mitigate overfitting. The decoder employs transposed convolutions for upsampling and incorporates skip connections to retain spatial details from the encoder layers.

The final layer applies a sigmoid activation function to produce a binary segmentation mask. This design balances computational efficiency with the ability to capture fine-grained details in the input data.

3.3. Training Strategy

The model is trained using the Adam optimizer with an initial learning rate of 0.01, adjusted adaptively using a ReduceLROnPlateau scheduler. The combination of Binary Cross-Entropy (BCE) Loss and Focal Tversky Loss is employed to address class imbalance, with BCE focusing on overall accuracy and Focal Tversky Loss emphasizing hard-to-segment regions.

4. Results

4.1. Training Progression

Over the course of 40 epochs, the model demonstrated stable convergence. The training loss consistently decreased, and the validation loss reached its minimum of 0.0650, reflecting effective learning and no significant overfitting.

4.2. Quantitative Results

The model achieved a final test Mean IoU of 63.24% at a threshold of 0.5, demonstrating its ability to accurately segment sparse foreground regions in highly imbalanced datasets.

4.3. Qualitative Results

Qualitative analysis revealed that the model effectively segmented sparse and irregular foreground regions while maintaining precise boundaries. The ResNet-inspired U-Net architecture and the use of Focal Tversky Loss contributed to these improvements, enabling the model to overcome challenges posed by the dataset's class imbalance and noise. Representative examples of input spectrograms, ground truth masks, and predicted masks are shown below. These results highlight the model's ability to identify sparse foreground regions in noisy conditions.

5. Discussion

This study effectively addresses class imbalance in segmentation tasks by employing:

- Residual connections to improve gradient flow and hierarchical feature representation.
- The Focal Tversky Loss to emphasize hard-to-segment regions, reducing background bias.
- Decoding strategies using transposed convolutions to recover spatial resolution.

Foreground composition analysis revealed that more than 90% of spectrograms have vocalization regions covering less than 10% of the total area. This imbalance significantly increases the risk of background dominance during training. Our architecture, combined with targeted loss functions, mitigates this challenge by balancing penalties for false positives and false negatives. Our model's IoU of 63.24% demonstrates the advantage of integrating residual connections and tailored loss functions. While this improvement is notable, further advancements are possible by incorporating data augmentation or higher-resolution spectrograms.

Future research should extend this framework to multi-class segmentation tasks and explore temporal features to

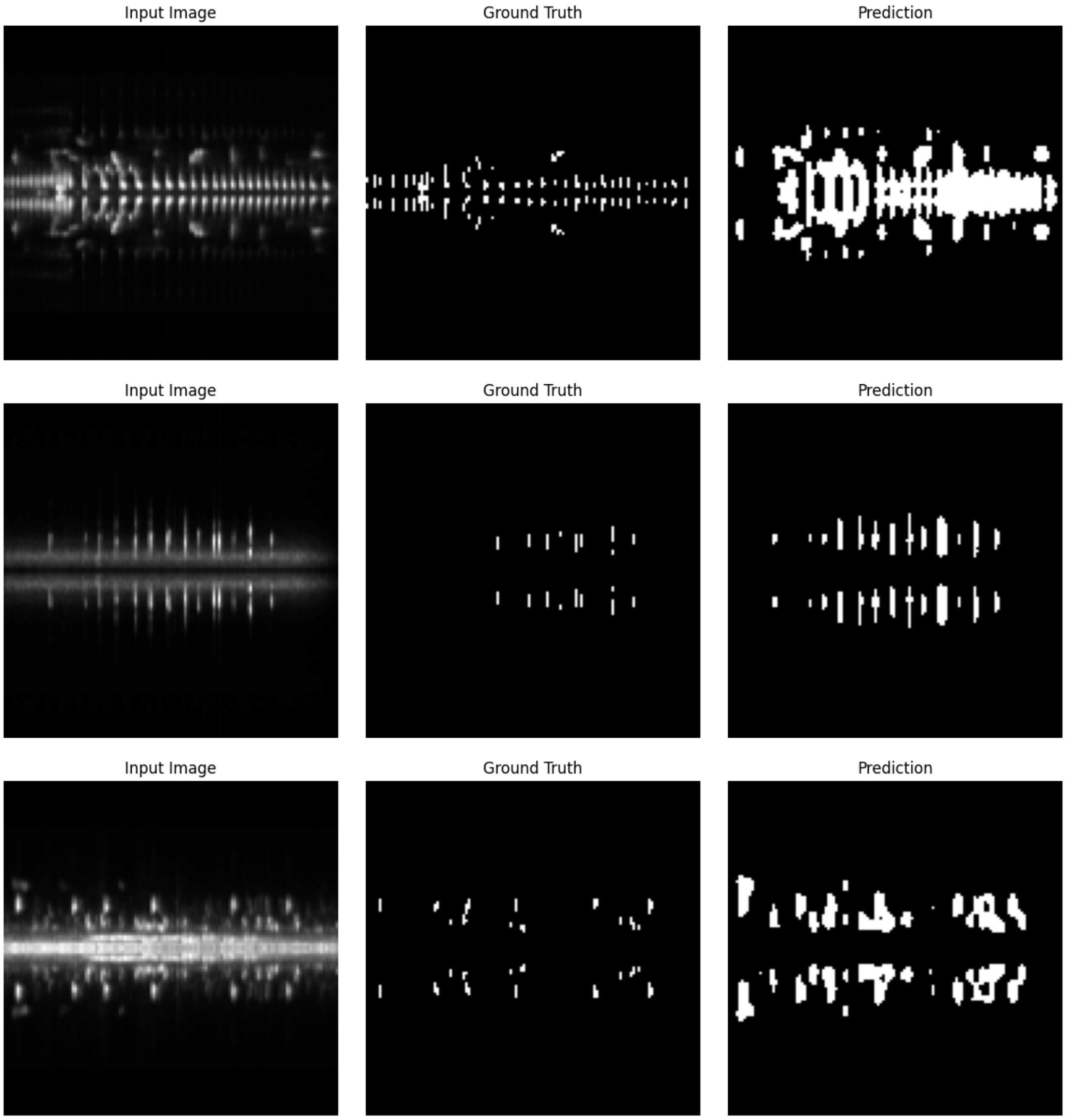

Figure 3. Qualitative results: Input spectrograms, ground truth masks, and predicted masks.

better capture vocalization dynamics over time. Additionally, lightweight architectures could enable deployment on edge devices for real-time ecological applications.

6. Conclusion

This paper introduces a ResNet-inspired U-Net architecture leveraging the Focal Tversky Loss to address segmentation challenges in imbalanced datasets. By combining hierarchical feature extraction, residual connections, and tailored loss functions, the model achieves a mean IoU of 63.24% on a bird sound segmentation dataset, demonstrating its robustness against foreground sparsity and noise.

These results highlight the significance of specialized architectures in bioacoustic analysis. Future work will explore incorporating temporal data, lightweight model designs for edge deployment, and techniques for augmenting sparse datasets to further enhance segmentation performance in ecological and conservation applications.

References

[1] Nabila Abraham and Naimul Mefraz Khan. A novel focal tversky loss function with improved attention u-net for lesion segmentation. *IEEE Transactions on Medical Imaging*, 39:378–388, 2019. 1, 2

[2] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully con-

nected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018. 2

[3] Sahil Kumar, Jialu Li, and Youshan Zhang. Vision transformer segmentation for visual bird sound denoising. *ArXiv preprint arXiv:2406.09167*, 2024. 2

[4] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 1, 2

[5] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen AW van der Laak, Bram van Ginneken, and Clara I S
'anchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017. 1

[6] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention*, 2015. 1, 2

[7] Seyed Sadegh Mohseni Salehi, Deniz Erdogmus, and Ali Gholipour. Tversky loss function for image segmentation using 3d fully convolutional deep networks. *International Workshop on Machine Learning in Medical Imaging*, 2017. 1, 2

[8] Prabhakar Tharun. Bird sound spectrogram segmentation using deep learning. *Bioacoustics Research Journal*, 2024. 1, 2

[9] Youshan Zhang and Jialu Li. Birdsoundsdenoising: Deep visual audio denoising for bird sounds. *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023. 1, 2