Optimizing Spectrogram Resolution and Training Strategies for Real-Time Killer Whale Call Type Classification

Vladislav Naumov[1] , **Iaroslav Sheipak**[1] , **Yuriy Ivanov**[2] and **Ilya Makarov**[3,4]
[1]MIRIAI
[2]HSE University
[3]AXXX
[4]Trusted AI Research Center, RAS
{vladnaumov2118, iaroslav.sheipak}@gmail.com, iuviivanov@edu.hse.ru, iamakarov@hse.ru

Abstract

Automated identification of killer whale call types from continuous acoustic recordings is essential for scalable population monitoring, yet existing general-purpose frameworks such as ANIMAL-SPOT suffer from suboptimal spectrogram resolution and lack strategies tailored to the spectral-temporal characteristics of killer whale vocalizations. We identify two key limitations of the ANIMAL-SPOT framework: (1) no possibility to choose different architecture of CNN backbone, and (2) the absence of regularization and class-balancing techniques limits generalization across call types with varying abundance. To address these issues, we propose a framework that combines optimized STFT parameters (FFT size 1024, hop length 172) with label smoothing and targeted oversampling, evaluated across three CNN backbones and five segment lengths. On a dataset of 12 killer whale vocalization classes from Avacha Gulf, Russia, our best configuration (ResNet-18 with 1200 ms segments) achieves **97.1%** segment-level accuracy, compared to **96.2%** for the ANIMAL-SPOT baseline — a relative error reduction of 23.7%. We further present an interactive demonstration system for uploading recordings and obtaining time-resolved call-type predictions, enabling rapid analysis of passive acoustic monitoring data.

1 Introduction

Killer whales (*Orcinus orca*) produce stereotyped, socially learned vocalizations that encode pod identity and cultural transmission [Ford, 1991; Yurk *et al.*, 2002; Filatova *et al.*, 2017; Deecke *et al.*, 2000]. Accurate call-type classification is essential for studying population dynamics and informing conservation, yet manual spectrogram inspection is labor-intensive and difficult to scale to modern passive acoustic datasets [Mellinger and et al., 2007; Gibb *et al.*, 2019]. Deep learning provides an efficient alternative by learning spectral–temporal structure directly from audio representations [Stowell *et al.*, 2015; Aodha *et al.*, 2018].

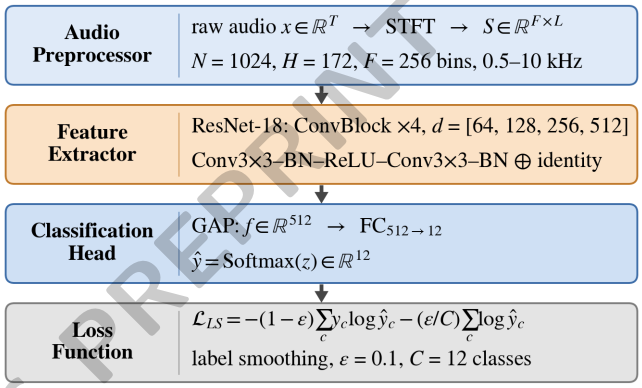

Figure 1: Overall architecture of the proposed model framework for call-type classification.

General-purpose frameworks such as ANIMAL-SPOT [Bergler *et al.*, 2022] are species-independent but suboptimal for killer whale call types. Their default STFT settings prioritize frequency over temporal resolution, limiting discrimination of short, frequency-modulated calls. They also lack regularization and class-balancing strategies, reducing robustness to uneven call distributions.

We optimize STFT parameters, segment length, and training strategy for killer whale call-type classification, and deploy the best model in an interactive demo for time-resolved predictions on continuous recordings. The proposed model architecture is illustrated in Figure 1.

The main contributions of this work are:

- A systematic analysis showing that STFT parameter optimization (FFT size, hop length) is the dominant factor in improving killer whale call type classification accuracy over the ANIMAL-SPOT baseline.
- An ablation study across segment lengths, backbones, and training strategies, demonstrating that task-specific signal processing choices outweigh architectural complexity.
- An interactive demonstration system for real-time call-type prediction from continuous recordings.

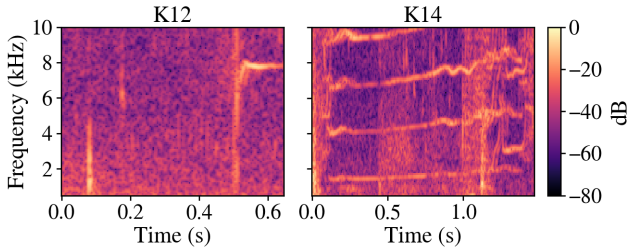

Figure 2: Spectrogram examples of K12 and K14 call types, illustrating their structural similarity that makes fine-grained classification challenging.

2 Related Work

Deep learning has increasingly replaced manual annotation in animal bioacoustics, with convolutional and recurrent networks applied to birds [Stowell *et al.*, 2015], bats [Aodha *et al.*, 2018], and marine mammals [Bergler *et al.*, 2022]. By learning spectral–temporal features directly from audio representations, these methods achieve greater accuracy and scalability than traditional feature-based approaches. For multi-class killer whale call classification, spectrogram design is critical, as call types often overlap in frequency but differ in temporal modulation; yet most frameworks rely on fixed STFT settings optimized for detection rather than fine-grained discrimination. ANIMAL-SPOT [Bergler *et al.*, 2022] serves as a strong, species-independent baseline, though its default preprocessing is not tailored to killer whale vocalizations. Domain-specific preprocessing can substantially improve performance. Label smoothing [Szegedy *et al.*, 2016] enhances generalization under ambiguous boundaries, and oversampling mitigates class imbalance.

3 Problem Statement

Given a continuous recording, the audio is divided into fixed-length segments and each segment is classified into one of C call types or noise. Let $x \in \mathbb{R}^T$ denote a segment of T samples. We compute a spectrogram $S \in \mathbb{R}^{F \times L}$ using the short-time Fourier transform, where F and L denote frequency bins and time frames. A classifier $f_\theta : \mathbb{R}^{F \times L} \to \{1, \ldots, C\}$ maps S to a predicted label.

Performance depends primarily on (1) STFT parameters (FFT size N, hop length H), which govern the time–frequency trade-off; (2) segment duration T, defining acoustic context; and (3) training strategy, including regularization and class balancing.

4 Proposed Framework

4.1 Spectrogram Computation

Audio segments are resampled to 44.1 kHz and converted to magnitude spectrograms using the STFT with FFT size $N = 1024$ and hop length $H = 172$. Frequency bins are restricted to the 500–10,000 Hz range to focus on the killer whale vocal range, and the resulting spectrogram is resized to 256 frequency bins via linear interpolation. Magnitudes are converted to decibels and normalized to $[0, 1]$.

4.2 Model Architecture

The backbone is a ResNet-18 [He *et al.*, 2016] with the first convolutional layer adapted for single-channel spectrogram input ($1\times7\times7$ convolution instead of $3\times7\times7$). The network processes the spectrogram through residual blocks, followed by global average pooling and a fully connected layer with C output nodes. We also evaluate ResNet-34 [He *et al.*, 2016] and ConvNeXt-Tiny [Liu *et al.*, 2022] as alternative backbones to assess whether increased model capacity or modern architectural design improves classification accuracy.

Label smoothing. To reduce overconfidence and improve generalization, we apply label smoothing [Szegedy *et al.*, 2016] with smoothing factor $\epsilon = 0.1$. This replaces hard one-hot targets with soft distributions, encouraging the model to learn more robust decision boundaries between acoustically similar call types.

Targeted oversampling. Although the dataset is approximately balanced across call types (∼3,000 samples per class), the noise class is substantially larger (∼12,000 samples). We apply weighted random sampling with weights proportional to the square root of the inverse class frequency, ensuring that all call types are represented equally across training epochs.

Optimization. All models are trained with AdamW, a one-cycle learning rate schedule (peak learning rate 3×10^{-3}), mixed-precision training (bfloat16), and gradient clipping at 1.0. Early stopping with patience of 25 epochs is used to prevent overfitting. The batch size is 64 for ResNet models and 32 for ConvNeXt-Tiny.

5 Experimental Setup

We use a dataset of field recordings of killer whale vocalizations collected in Avacha Gulf, Russia. We sincerely thank Olga Filatova for providing the dataset and for her valuable insights in interpreting the results. The dataset contains 45,429 labeled audio segments spanning 11 stereotyped call types (K1, K4, K5, K7, K10, K12, K13, K14, K17, K21, K27) and a noise class, for a total of 12 classes. Call type classes contain approximately 3,000 segments each, while the noise class contains 12,228 segments. Data are split on recording tape to prevent information leakage, resulting in 70%/ 15% / 15% splits between training / validation / tests. Among the call types, K12 and K14 (Figure 2) are particularly challenging because they share similar frequency contours and differ primarily in subtle temporal modulations, making high temporal resolution essential for their discrimination.

Segment-level classification accuracy is adopted as the primary metric. Per-class accuracy is reported for K12 and K14 to evaluate performance on the most challenging call types. All configurations are evaluated on the same held-out test set. We also report end-to-end real-time factor (RTF), defined as total wall-clock time (I/O, preprocessing, inference, and post-processing) divided by input audio duration.

As baseline, we use ANIMAL-SPOT with the ResNet-18 architecture and STFT parameters reported in prior work on this dataset: FFT size 1024, hop length 300, 48 kHz sampling rate, frequency range 50–10,000 Hz, and segment

Backbone	Seq (ms)	Label smooth.	FFT / Hop	Over-samp.	K12	K14	Overall
ResNet-18†	800	✗	1024 / 300	✗	**0.928**	0.933	0.962
ResNet-18	800	✓	1024 / 172	✗	0.800	0.936	0.960
	800	✓	1024 / 172	✓	0.882	0.936	0.967
	1000	✓	1024 / 172	✓	0.846	**1.000**	0.967
	1200	✓	1024 / 172	✓	0.894	0.936	**0.971**
	1500	✓	1024 / 172	✓	0.831	0.936	0.961
ResNet-34	800	✓	1024 / 172	✓	0.856	0.936	0.969
ConvNeXt-Tiny	1000	✓	1024 / 172	✓	0.826	**1.000**	0.953
	1200	✓	1024 / 172	✓	0.840	0.873	0.948

Table 1: Ablation study of model configurations on the test set. The baseline (top row, shaded) and our best configuration are highlighted. †Baseline uses 48 kHz sampling rate and fmin = 50 Hz. Bold indicates best per-column values.

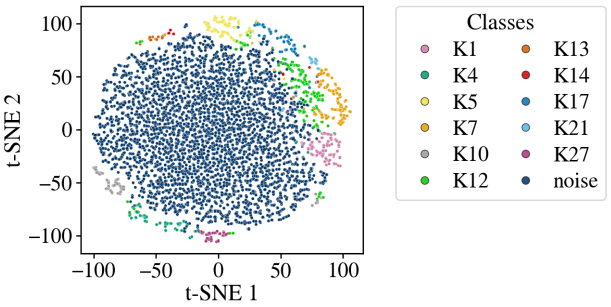

Figure 3: t-SNE visualization of segment embeddings from the penultimate layer of the best model, showing class separability across killer whale call types.

length 800 ms. No label smoothing or oversampling is applied. This configuration achieves 96.2% accuracy and serves as the reference for our ablation study.

6 Results

Table 1 presents the results of our ablation study. The ANIMAL-SPOT baseline (FFT 1024, hop 300, 800 ms at 48 kHz) achieves 96.2% overall accuracy without label smoothing or oversampling. The key findings are as follows.

Optimized STFT and training strategies provide consistent gains. Combining a finer hop length (172 vs 300), longer segments (1200 ms vs 800 ms), label smoothing, and targeted oversampling improves overall accuracy from 96.2% to 97.1% (+0.9 pp), representing a 23.7% relative error reduction. This demonstrates that jointly optimizing signal processing and training strategies yields meaningful gains even when the baseline uses a reasonable STFT configuration.

Segment length has a clear optimum. Within our training configuration, increasing segment length from 800 ms to 1200 ms improves accuracy from 96.7% to 97.1%, as longer segments capture more acoustic context. However, further increasing to 1500 ms reduces accuracy to 96.1%, indicating that excessively long segments introduce noise and dilute discriminative features. The optimal segment length of 1200 ms balances context and precision.

Oversampling improves rare class accuracy. At 800 ms segment length, adding targeted oversampling improves K12 accuracy from 80.0% to 88.2% (+8.2 pp) and overall accuracy from 96.0% to 96.7%, demonstrating its effectiveness for balancing class representation during training.

Backbone architecture has limited impact. ResNet-34 (96.9%) and ConvNeXt-Tiny (95.3% at 1000 ms, 94.8% at 1200 ms) do not outperform ResNet-18 (97.1%), indicating that the spectral-temporal patterns in killer whale calls are adequately captured by a lightweight backbone. The additional capacity of deeper or more modern architectures does not translate to improved classification.

7 Discussion

The t-SNE projection of segment embeddings [van der Maaten and Hinton, 2008] (Figure 3) shows well-separated clusters for most call types, with partial K12–K14 overlap consistent with their structural similarity; the noise class is clearly separated.[1]

These findings indicate that, for fine-grained acoustic classification, spectrogram parameterization—particularly the STFT trade-off between temporal and frequency resolution—can influence performance more strongly than network architecture. This insight likely generalizes to other tasks involving short, temporally modulated signals. A limitation is the use of a single regional dataset; future work should evaluate cross-population generalization across dialects and recording conditions.

8 Demonstration System

We additionally present an interactive web-based demonstration system that enables researchers to upload continuous acoustic recordings and obtain time-resolved call-type predictions. The system segments the input audio into overlapping windows, classifies each segment using the trained model, and outputs predictions in Raven selection table format, a standard interchange format for bioacoustic annotation tools. This enables rapid exploration of large passive acoustic monitoring datasets without requiring expertise in deep learning, and provides a practical tool for field researchers studying killer whale vocal behavior. On an RTX 5090, our best ResNet-18 model achieves end-to-end RTF=0.007.

9 Conclusion

We presented a systematic study of spectrogram resolution and training strategies for killer whale call type classification, achieving 97.1% segment-level accuracy across 12 classes. Through ablation, we demonstrated that label smoothing and targeted oversampling improve accuracy by +0.9 pp over the unregularized baseline, with segment length optimization at 1200 ms providing the best trade-off between acoustic context and classification precision. Backbone architecture had negligible impact, with the lightweight ResNet-18 outperforming both ResNet-34 and ConvNeXt-Tiny. The accompanying interactive demonstration system enables practical, real-time analysis of field recordings in standard bioacoustic formats.

[1] Supplementary material available online: demo video (link) and presentation slides (link).

Acknowledgments

The work was supported by a grant, provided by the Ministry of Economic Development of the Russian Federation (agreement dated June 20, 2025 No. 139-15-2025-011, identifier 000000C313925P4G0002).

References

[Aodha *et al.*, 2018] Dustin Mac Aodha, Roger Gibb, Kevin Barlow, and et al. Bat detective—deep learning tools for bat acoustic signal detection. *PLOS Computational Biology*, 14(3):e1005995, 2018.

[Bergler *et al.*, 2022] Christian Bergler, Felipe Tenorio, Yannis Stylianou, and et al. Animal-spot: An open, generalizable deep learning framework for animal sound detection and classification. *Scientific Reports*, 12:21485, 2022.

[Deecke *et al.*, 2000] Volker B. Deecke, John K. B. Ford, and Paul Spong. Dialect change in resident killer whales: implications for vocal learning and cultural transmission. *Animal Behaviour*, 60:629–638, 2000.

[Filatova *et al.*, 2017] Olga A. Filatova, Tatiana V. Ivkovich, Mikhail A. Guzeev, Alexander M. Burdin, and Erich Hoyt. Social complexity and cultural transmission of dialects in killer whales. *Behaviour*, 154(2):171–194, 2017.

[Ford, 1991] John K. Ford. Vocal traditions among resident killer whales (Orcinus orca) in coastal waters of british columbia. *Canadian Journal of Zoology*, 69(6):1454–1483, 1991.

[Gibb *et al.*, 2019] Rory Gibb, Dustin Mac Aodha, and et al. Emerging opportunities and challenges for passive acoustics in ecological assessment and monitoring. *Methods in Ecology and Evolution*, 10(2):169–185, 2019.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[Liu *et al.*, 2022] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, and et al. A convnet for the 2020s. *arXiv preprint arXiv:2201.03545*, 2022.

[Mellinger and et al., 2007] David K. Mellinger and et al. An overview of fixed passive acoustic observation methods for cetaceans. *Oceanography*, 20(4):36–45, 2007.

[Stowell *et al.*, 2015] Dan Stowell, Mark D. Wood, Yannis Stylianou, and Hervé Glotin. Automatic acoustic detection of birds through deep learning: the first bird audio detection challenge. *Methods in Ecology and Evolution*, 6(8):977–986, 2015.

[Szegedy *et al.*, 2016] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016.

[van der Maaten and Hinton, 2008] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[Yurk *et al.*, 2002] Heidi Yurk, Lance Barrett-Lennard, John K. B. Ford, and Craig O. Matkin. Cultural transmission within maternal lineages: vocal clans in resident killer whales in southern alaska. *Animal Behaviour*, 63:1103–1119, 2002.