TRANSFORMER OBJECT DETECTION ON SPECTROGRAMS FOR ANTARCTIC BLUE AND FIN WHALE CALL DETECTION

Technical Report

Xixin Zhang

San Diego State University, San Diego, CA, USA
xzhang3906@sdsu.edu

ABSTRACT

We cast the detection of strongly-labelled Antarctic baleen whale calls in the BioDCASE-2026 Task 2 data as a two-dimensional object-detection problem on spectrogram images, and address it with RF-DETR, a detection transformer built on a self-supervised DINOv2 backbone. Calls are detected as bounding boxes in the time–frequency plane and projected onto the time axis, and the predictions from overlapping spectrogram windows are consolidated by confidence thresholding and temporal non-maximum suppression. On the development validation data the system attains an F_1 of 0.584, compared with 0.438 for the official YOLOv11 baseline under the same evaluation protocol, a relative improvement of about 33%.

***Index Terms*—** sound event detection, bioacoustics, baleen whale, detection transformer, transfer learning

1. INTRODUCTION

BioDCASE 2026 Task 2 [1], the second edition of the BioDCASE whale-call detection task [2], asks for the detection and classification of seven call types produced by two Antarctic baleen whale species: the blue whale (*Balaeonoptera musculus*) (BmA, BmB, BmZ, BmD) and the fin whale (*Balaenoptera physalus*) (BpD, Bp20, Bp20+). For scoring, these seven types are collapsed into three groups—the blue-whale ABZ call (BmA/BmB/BmZ), the downsweep (BmD/BpD), and the fin-whale 20 Hz pulse (Bp20/Bp20+)—and performance is measured by per-class Precision, Recall and F_1, where a detection is counted as correct when its temporal intersection-over-union (IoU) with a reference event exceeds 0.3 [3].

The official baseline [3] renders each recording as a sequence of spectrograms and detects calls with a YOLOv11s object detector, exploiting the fact that whale calls have a characteristic, visually distinctive footprint in the time–frequency plane. We keep this detection-on-spectrograms formulation but replace the convolutional detector with a detection transformer initialised from a strong visual foundation model, which we find transfers well to the spectrogram domain despite the large gap from its natural-image pretraining.

2. METHOD

2.1. Time–frequency representation

Each recording is processed at a sampling rate of 250 Hz, which covers the full bandwidth of the target calls (below 125 Hz). A recording is divided into 80 s windows with 50% overlap, so that a call near a window boundary is captured intact in at least one window. Every window is transformed by a short-time Fourier transform with a 2.05 s analysis window and a 0.31 s hop; the resulting magnitude is mapped to decibels and rendered, with low frequencies at the bottom, as a single-channel image of 257×260 pixels (frequency $\times$ time). Reference annotations are expressed as bounding boxes in window-local time–frequency coordinates.

2.2. Detector

The detector is RF-DETR Large [4], a DETR-style [5] architecture whose backbone is a DINOv2 [6] ViT-L encoder pre-trained by self-supervision on a large natural-image corpus. The detection transformer predicts a fixed set of boxes directly, without anchor boxes or a hand-tuned suppression heuristic in the detector itself, which suits the sparse and variable-duration nature of whale calls. The single-channel spectrogram is replicated to three channels to match the backbone; the classification head is sized for the seven call types and the architecture is otherwise unchanged.

2.3. Training

The detector is fine-tuned for 20 epochs on the development training split with the optimisation settings released with RF-DETR (AdamW, base learning rate 10^{-4} and backbone learning rate 10^{-5}, weight decay 10^{-3}, batch size 16, exponential-moving-average of the weights with decay 0.993) on a single NVIDIA A100 GPU. Spatial and photometric augmentations (random crop, horizontal flip) are applied to the spectrogram images. We retain the exponential-moving-average weights that give the best validation F_1, which occur at epoch 10. The DINOv2 backbone is the only external resource used; no external audio is employed.

2.4. Inference and post-processing

The transformer emits many overlapping candidate boxes, both within a window and across the 50% window overlap. We first discard candidates whose confidence falls below a threshold τ, then apply temporal non-maximum suppression: within each recording and predicted class, candidates are considered in order of decreasing confidence and any candidate whose temporal IoU with an already-retained one exceeds 0.3 is removed. The surviving detections are mapped from the seven call types to the three evaluation groups, and their window-local times are converted back to absolute time. We set $\tau = 0.4$ based on the sweep reported in Section 3 (Table 3).

Table 1: Overall performance on the development validation split. Precision and Recall are averaged across the three evaluation groups; F_1 is derived from the averaged values.

System	Precision	Recall	F_1
YOLOv11 baseline [3]	0.498	0.390	0.438
RF-DETR (this work)	**0.555**	**0.616**	**0.584**

Table 2: Per-deployment performance of the proposed system on the validation split.

Deployment	Precision	Recall	F_1
casey2017	0.558	0.657	0.603
kerguelen2014	0.592	0.558	0.574
kerguelen2015	0.578	0.675	0.622

3. EXPERIMENTS

3.1. Protocol

We evaluate on the development validation split, which comprises 587 one-hour recordings drawn from three deployments (casey2017, kerguelen2014, kerguelen2015), using the official evaluation procedure at the temporal IoU threshold of 0.3 [3]. Following the challenge convention, per-class Precision and Recall are averaged with equal weight across the three evaluation groups, and F_1 is computed from the averaged Precision and Recall. As a reference point we reproduce the official YOLOv11 baseline from its released checkpoint [7] under the identical protocol.

3.2. Results

Table 1 reports the overall comparison. The proposed system raises F_1 from 0.438 to 0.584, improving both Precision and Recall over the baseline rather than trading one for the other. Table 2 breaks the result down by deployment; performance is consistent across the three sites, with F_1 between 0.57 and 0.62.

3.3. Effect of post-processing

Table 3 sweeps the confidence threshold τ from 0.1 to 0.5 in steps of 0.1, with and without temporal non-maximum suppression (NMS). Two effects are clear. First, NMS helps at every threshold, with the largest gains at low τ where the overlap-induced duplicates are most numerous (e.g. $+0.13$ F_1 at $\tau = 0.2$). Second, F_1 rises with τ up to 0.4 and then declines, so $\tau = 0.4$ with NMS is the operating point we submit. The two stages are complementary: thresholding alone plateaus at 0.52, and NMS lifts the peak to 0.584.

3.4. Per-class results

Table 4 gives the per-group breakdown at the submitted operating point. The fin-whale 20 Hz pulse is detected most precisely ($F_1 = 0.63$) and the blue-whale ABZ call has the highest Recall (0.81), while the downsweep is the hardest group ($F_1 = 0.48$): it is the rarest and most easily confused class, and brings down the averaged score.

Table 3: Validation F_1 as a function of the confidence threshold τ, with and without temporal NMS. Best in bold.

τ	no NMS	with NMS
0.1	0.112	0.222
0.2	0.287	0.420
0.3	0.432	0.535
0.4	0.519	**0.584**
0.5	0.519	0.551

Table 4: Per-class performance at the submitted operating point ($\tau = 0.4$ with NMS) on the validation split.

Call group	Precision	Recall	F_1
ABZ call (BmA/BmB/BmZ)	0.474	0.805	0.597
Downsweep (BmD/BpD)	0.471	0.488	0.480
Fin 20 Hz (Bp20/Bp20+)	0.721	0.555	0.627

4. DISCUSSION

The result indicates that a detection transformer with a strong visual prior is a competitive backbone for bioacoustic event detection, even though its pre-training is on natural images rather than spectrograms. Several directions could narrow the remaining gap. An audio-pre-trained encoder such as BEATs [8] or Audio-MAE [9] would match the input domain more closely and might transfer better than a vision backbone. Our single confidence threshold is shared across all call types; tuning it per evaluation group would let the system trade Precision against Recall where each call type benefits most. The detector also receives no explicit prior on the frequency band that each call type occupies, information that a purpose-built acoustic detector would exploit. Finally, the reported figures come from a single model without ensembling.

A caveat for the challenge evaluation is that the development and evaluation data come from disjoint deployments: the validation recordings are from casey2017, kerguelen2014 and kerguelen2015, whereas the evaluation set is drawn from ddu2019, ddu2021, kerguelen2019 and kerguelen2020. The shift in site and year introduces a soundscape domain gap, so some drop relative to the validation numbers reported here is expected on the evaluation set.

5. REFERENCES

[1] BioDCASE 2026 Challenge, "Task 2: Supervised detection of strongly-labelled whale calls," https://biodcase.github.io/challenge2026/task2, 2026.

[2] D. Stowell, E. Vidaña Vila, I. Nolasco, B. McEwen, L. Jean-Labadye, Y. Benhamadi, G. Dubus, B. Hoffman, P. Linhart, I. Morandi, D. Cazau, B. Miller, E. Schall, C. Parcerisas, A. Gros-Martial, I. Moummad, P.-Y. Raumer, E. White, P. White, P. Nguyen Hong Duc, and V. Lostanlen, "BioDCASE: Using data challenges to make community advances in computational bioacoustics," *bioRxiv*, 2026, doi:10.64898/2026.04.02.716062.

[3] C. Parcerisas, "YOLOv11 baseline for the BioDCASE 2025 task 2 challenge," GitHub repository, https://github.com/marinebioCASE/task2_2025, 2025.

[4] I. Robinson, P. Robicheaux, M. Popov, D. Ramanan, and N. Peri, "RF-DETR: Neural architecture search for real-time detection transformers," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2026, arXiv:2511.09554.

[5] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. European Conf. Computer Vision (ECCV)*, 2020, pp. 213–229.

[6] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "DINOv2: Learning robust visual features without supervision," *Trans. Machine Learning Research*, 2024, arXiv:2304.07193.

[7] C. Parcerisas, "Baseline model weights for task 2 BioDCASE," Zenodo, https://doi.org/10.5281/zenodo.20281116, 2026.

[8] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, W. Che, X. Yu, and F. Wei, "BEATs: Audio pre-training with acoustic tokenizers," in *Proc. Int. Conf. Machine Learning (ICML)*, 2023, pp. 5178–5193.

[9] P.-Y. Huang, H. Xu, J. Li, A. Baevski, M. Auli, W. Galuba, F. Metze, and C. Feichtenhofer, "Masked autoencoders that listen," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.