DRAWING BOXES IN THE DEEP: SPECTROGRAM OBJECT DETECTION FOR ANTARCTIC WHALE CALLS

Technical Report

Michael Moshe Michelashvili, Danielle Hausler, Amit Galor,
Shai Nahum Gefen, Tomer Nachshon, Yuval Mendelson, Naama Yochai

Deep Voice Foundation

{moshe, danielle, amit, shai, tomer, yuval, naama}@deepvoicefoundation.com

ABSTRACT

We present a detection system for strongly-labelled Antarctic blue and fin whale calls across three classes (`bmabz`, `d`, `bp`) in long-duration 250 Hz recordings. In contrast to frame-level classifiers, our primary approach casts sound-event detection as two-dimensional object detection on spectrogram images. The input audio is windowed into 30 second chunks with 50% overlap and rendered as an inverted, 98th-percentile-normalised magnitude spectrogram (512-point FFT, 20-sample hop), which is processed by a YOLOv11m object detector. Predicted bounding boxes are projected onto the time axis and mapped to absolute timestamps via each recording's start time. Detections from overlapping windows are merged by per-class temporal non-maximum suppression, followed by span dilation, physical-duration capping, minimum-duration filtering, and per-class confidence thresholding. We further train three detectors on complementary spectrogram representations (per-bin median-subtracted, mean-plus-median-subtracted, and grayscale magnitude) and route each class to the representation best suited to it: `bmabz` from median-subtracted, `d` from mean-plus-median-subtracted, and `bp` from grayscale. Trained on the provided training set and evaluated on the validation set, the single detector yields an overall F1 of 0.612, and the three-stream per-class fusion 0.635. Four systems are submitted: the single detector (`task2_1`); the three-stream fusion (`task2_2`), our strongest system; a recall-oriented ensemble (`task2_3`) that adds non-overlapping detections from an independently designed, MAE-pretrained Vision Transformer frame-level detector operating on 16 second chunks; and that Vision Transformer detector in isolation (`task2_4`), included as an architecturally independent alternative. Among these, the fusion favours precision while the gap-fill ensemble favours recall.

***Index Terms*—** Sound event detection, bioacoustics, whale call detection, object detection, vision transformer

1. INTRODUCTION

BioDCASE 2026 Task 2 requires temporal detection of Antarctic blue and fin whale vocalisations in long passive-acoustic recordings from the International Whaling Commission Southern Ocean Research Partnership (IWC-SORP) Acoustic Trends project [1]. We developed two independent detector families. The first (Section 3) treats the spectrogram as an image and uses a YOLOv11m object detector (You Only Look Once, YOLO); it underlies **YOLO** and **YOLO-3×**. The second (Section 4) is a Vision Transformer (ViT) pretrained as a masked autoencoder (MAE), predicting per-frame class probabilities; combined with YOLO-3× it forms **YOLO+ViT**. Section 2 describes the task and data, Section 5 the four submitted systems, and Section 6 the validation results.

2. TASK AND DATA

The development set is drawn from the Antarctic Blue and Fin whale Library and split into eight training deployments (∼6,000 one-hour files) and three validation deployments (`casey2017`, `kerguelen2014`, `kerguelen2015`; ∼590 files). The evaluation set comprises four further site-years (`ddu2019`, `ddu2021`, `kerguelen2019`, `kerguelen2020`) released without public annotations [2]. Three classes are scored: `bmabz` (blue-whale A/B/Z calls), `d` (D-call / downsweep, from the `bmd` and `bpd` annotations), and `bp` (fin-whale 20 Hz pulse). Each system must output a comma-separated values (CSV) file of `(dataset, filename, annotation, start_datetime, end_datetime)`. Scoring uses a one-dimensional temporal intersection-over-union (IoU > 0.3) between predicted and ground-truth intervals; critically, when several predictions overlap one ground-truth event, only one is counted as a true positive and the rest as false positives, so an accurate *count* of calls matters as much as coverage. Precision, recall and F1 are computed per class and per deployment.

3. YOLO SPECTROGRAM-DETECTION SYSTEMS

This family underlies **YOLO** (`task2_1`) and **YOLO-3×** (`task2_2`).

3.1. Acoustic representation

Recordings are processed at their native 250 Hz rate, split into 30 s chunks with 50% overlap, band-limited (5-124 Hz) and transformed with a magnitude short-time Fourier transform (STFT, $\texttt{nfft} = 512$, hop $= 20$). The magnitude image is inverted, normalised by its 98th percentile, and written as a 320 px grayscale image. For the fusion system we additionally compute two background-suppressed variants of the same STFT, a per-frequency-bin *median*-subtracted and a per-bin *mean*-subtracted image, which suppress stationary background noise and expose different call types to different degrees (Fig. 1).

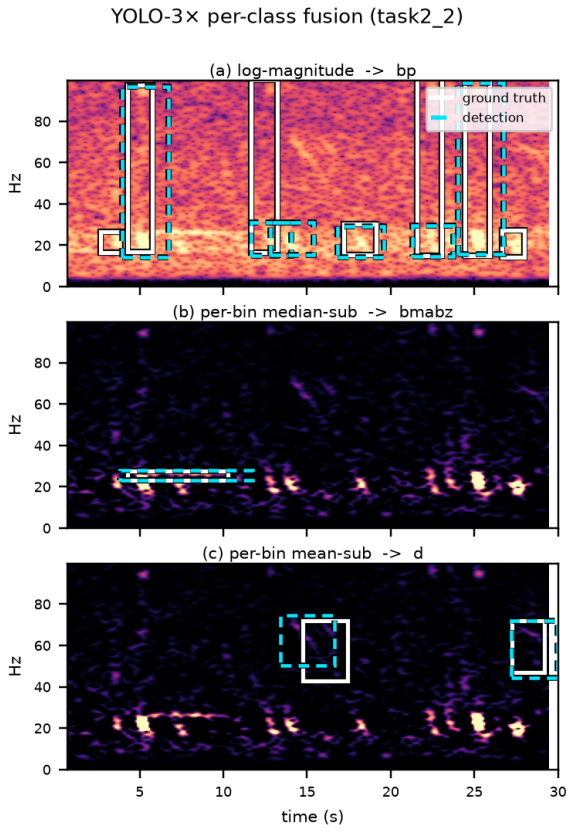

Figure 1: YOLO-3× on a validation window with all three classes. Each panel is one representation the fusion routes over and shows the class taken from it: (a) log-magnitude → `bp`, (b) per-bin median-subtracted → `bmabz`, (c) per-bin mean+median-subtracted → `d`. Per-bin subtraction exposes the calls; boxes are ground truth (white) and fusion detections (cyan).

3.2. Detector and training

Each stream is a YOLOv11m detector [3] initialised from COCO (Common Objects in Context) pretrained weights and fine-tuned on the eight development training deployments. Predicted boxes are projected onto the time axis (box centre $\pm$ width$/2$) and mapped to absolute timestamps via the start time encoded in each WAV filename. Training uses AdamW (`lr0` $= 10^{-3}$, one-epoch warmup, cosine schedule, 25 epochs, `imgsz` $= 320$) with class-balanced (1:1) background sampling. Augmentation is restricted to *spectrogram-safe* transforms only: hue-saturation-value (HSV) value-channel jitter, small translation and scale, and random erasing; flips, rotation and mosaic are disabled because the time and frequency axes are not interchangeable. This restriction was the single largest training lever ($+0.036$ F1 over default augmentation).

3.3. Per-class fusion (YOLO-3×)

YOLO-3× trains one detector per representation and takes each class from its best source: `bmabz` from the median-subtracted model, `d` from the mean+median-subtracted model, and `bp` from the grayscale model. The median-subtraction channel cleanly improves `d` at no cost to the others, and per-class routing banks each representation's strength.

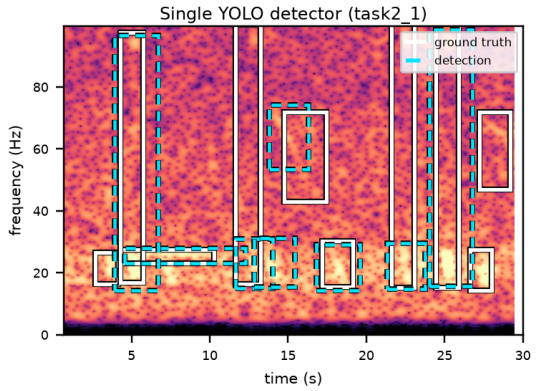

Figure 2: Single YOLO detector (`task2_1`) on the same window: detections (cyan) versus ground truth (white) on the log-magnitude spectrogram, all three classes from one detector.

3.4. Inference and post-processing

Per-window detections are merged and filtered per class, with weights fixed. We report two validation views (defined in Section 6): a *pooled* F1 and a *cross-site* F1. **Base post-processing** (both YOLO systems): temporal non-maximum suppression to remove overlap-window duplicates (per-class IoU `bmabz`/`d` $= 0.3$, `bp` $= 0.5$); span dilation of $+0.5$ s per side for `bp` and `bmabz`; physical span caps (`bmabz` 40 s, `d` 12 s, `bp` 9 s); and per-class confidence thresholds (`bmabz` 0.35, `d` 0.25, `bp` 0.20). **Refinement** (YOLO-3× only): a minimum-duration filter dropping `d` detections shorter than 1.6 s, and overlap suppression keeping only the higher-confidence of two same-class detections overlapping by ≥ 0.8 of the shorter one's duration. We set this overlap fraction to 0.8 because smaller values (e.g. 0.6) raised the pooled F1 but lowered the cross-site F1; 0.8 improves both (Table 3).

4. VIT FRAME-LEVEL DETECTOR

This detector is independent of the YOLO family and underlies the YOLO+ViT ensemble. It predicts per-frame class probabilities rather than boxes.

4.1. Representation

The input is a 16 s chunk at 250 Hz. A log-power STFT (`nfft` $=$ 512, hop $= 64$, 256 frequency bins over 0-125 Hz) is augmented with a *temporal-gradient* channel - the frame-to-frame difference d/dt of the log-power spectrogram - which gives the model an explicit frequency-sweep cue (a downsweep shows negative gradients at high bins and positive at low bins). The two-channel input is patchified on a 16×31 grid of 32-dimensional patches (496 tokens).

4.2. MAE self-supervised pretraining

The backbone is a 12-layer ViT (384-dimensional, 6 heads, feed-forward dimension 1536; the full detector has $\sim$23 M parameters) pretrained as a Masked Autoencoder [5, 4] on $\sim$1,880 h of *unlabelled in-domain development audio* (no external corpus), with 75% random patch masking and a lightweight 4-layer decoder ($\sim$18 h

on a single GPU); the detector is then fine-tuned on the labelled training deployments. Pretraining uses 60 s chunks; the positional embeddings are bilinearly interpolated to the 16 s detector grid at fine-tuning. Self-supervised pretraining improved training stability and reduced overfitting relative to training the encoder from scratch.

4.3. Detector head and training

A learned, domain-initialised frequency-attention module weights bins per class, followed by *per-class* multi-scale dilated temporal-convolution heads whose receptive fields match each call's duration (`bmabz` kernel 3, dilation 1/2/4, ≈3.75 s; `bp` kernel 5, ≈6.25 s; `d` kernel 7, dilation 1/3/9, ≈13.75 s), producing 64 frames per chunk (0.25 s resolution). Training uses focal loss (γ=2, α=0.25), mixup (α=0.3) and gain jitter, stochastic depth (0.20), layer-wise learning-rate decay (0.75), 3× positive oversampling, boundary label smoothing, and exponential moving average (EMA) weights (0.999). A duration-aware loss penalises isolated high-probability frames shorter than each class's minimum call duration (`bmabz` 1 s, `bp` 0.5 s, `d` 2 s), suppressing spurious single-frame spikes.

5. SUBMITTED SYSTEMS

- **YOLO** (`Yochai_DV_task2_1`): a single YOLOv11m detector for all three classes with base post-processing. ∼20.1 M parameters.
- **YOLO-3×** (`Yochai_DV_task2_2`, our best): the three-stream per-class fusion of Section 3.3 with base post-processing and refinement. ∼60.2 M parameters.
- **YOLO+ViT** (`Yochai_DV_task2_3`): YOLO-3× augmented with the ViT detector of Section 4. ViT events that do *not* overlap any same-class YOLO event (max overlap 0.3 s) and clear a per-class confidence floor (`bmabz` 0.5, `d`/`bp` 0.3) are *added*; the ensemble only ever adds detections. ∼83 M parameters total. A deliberate recall-oriented hedge (Section 7).
- **ViT** (`Yochai_DV_task2_4`): the ViT detector of Section 4 on its own, decoded to events by per-class frame-probability thresholds (`bmabz` 0.40, `d`/`bp` 0.30) with per-class gap-merging and minimum-duration filtering. ∼23 M parameters; a single-model alternative to the YOLO family.

6. RESULTS

We evaluate on the three validation deployments with the official scorer at the fixed per-class thresholds used for submission. We report two views: the *pooled* F1 (Table 1; true/false positives aggregated over all validation deployments, then class-averaged precision and recall combined into F1), and a *cross-site* F1 (mean per-deployment F1 under leave-one-deployment-out threshold selection), a proxy for generalisation to the unseen evaluation sites.

YOLO-3× is the strongest system on F1 and precision, on every deployment (Table 1) and on every class (Table 2); YOLO+ViT attains the highest recall (0.627). YOLO-3× is our primary submission. Table 3 traces the path from the official YOLOv11 baseline (F1 0.43, reported by the challenge [1]) to our best system: spectrogram-safe augmentation is the single biggest training lever (plain YOLOv11m 0.555 to 0.591), base post-processing lifts the single detector to 0.612, per-class fusion adds most of the remaining gain (and most of the cross-site robustness) to 0.623, and the refinement step brings it to 0.635. Figs. 2 and 1 contrast the single detector and the fusion against ground truth on a validation window. YOLO+ViT attains high recall but at lower precision, netting below YOLO-3× overall (Section 7). The standalone ViT (`task2_4`) reaches 0.538: above the 0.43 baseline but below the YOLO systems, reflecting that its frame-level strength does not fully translate to event-level localisation.

Table 1: Validation results: pooled overall F1/precision/recall and per-deployment F1 (official scorer, fixed thresholds). Best per column in **bold**.

Model	F1	P	R	casey17	kerg14	kerg15
YOLO	0.612	0.608	0.616	0.552	0.590	0.652
YOLO-3×	**0.635**	**0.649**	0.623	**0.587**	**0.594**	**0.680**
YOLO+ViT	0.598	0.571	**0.627**	0.547	0.557	0.659
ViT	0.538	0.621	0.474	0.558	0.490	0.588

Table 2: Per-class validation F1 (pooled). Best per column in **bold**.

Model	`bmabz`	`d`	`bp`
YOLO	0.630	0.547	0.649
YOLO-3×	**0.635**	**0.604**	**0.651**
YOLO+ViT	0.628	0.527	0.630
ViT	0.570	0.468	0.518

Table 3: Path from the official YOLOv11 baseline to our best system (validation F1; pooled / cross-site). Cross-site is not computed for the rows above base post-processing (n/a).

Configuration	pooled	cross-site
Official YOLOv11 baseline	0.43	n/a
Plain YOLOv11m (default augmentation)	0.555	n/a
+ spectrogram-safe augmentation	0.591	n/a
+ base post-processing	0.612	0.585
+ per-class fusion (3 streams)	0.623	0.604
+ refinement (**YOLO-3×**)	**0.635**	**0.611**

7. DISCUSSION AND LIMITATIONS

The `d` (D-call) class is the hardest for both families. An audit of high-confidence `d` false positives found that roughly 78 of 80 had zero ground-truth overlap yet were morphologically real, unlabelled downsweeps, suggesting the class ceiling is partly a label-noise effect rather than a model gap. This motivates YOLO+ViT: because the ViT and YOLO families make largely independent errors, ViT detections in YOLO's silent regions are good candidates for genuinely missed events. On the under-labelled validation set this trades away precision and nets below YOLO-3×, but the evaluation set undergoes a more complete two-phase re-annotation; if those real-but-unlabelled events are labelled there, the added detections convert from false to true positives. YOLO-3× remains our primary recommendation and YOLO+ViT the recall-leaning alternative. The evaluation set has no public ground truth, so evaluation predictions could only be checked visually across all four deployments.

8. CONCLUSION

Casting whale-call detection as spectrogram object detection, with spectrogram-safe augmentation, per-class multi-representation routing, and per-class temporal post-processing, yields a strong and interpretable system (YOLO-3×, 0.635 validation F1). An independent MAE-pretrained ViT frame detector, combined by recall-only gap-filling, provides a complementary hedge whose value depends on the completeness of the evaluation annotations. Our code is available at `https://github.com/deep-voice/biodcase-2026`.

9. ACKNOWLEDGMENT

This work was supported by the Deep Voice Foundation's Marine Bioacoustics Initiative. We thank the BioDCASE organisers and the data providers at Flanders Marine Institute, Sorbonne University, the Alfred-Wegener Institute, the Australian Antarctic Division, and the Muséum national d'Histoire naturelle.

10. REFERENCES

[1] "BioDCASE 2026 Task 2: detection of Antarctic blue and fin whale calls," 2026. [Online]. Available: `https://biodcase.github.io/challenge2026/task2`

[2] BioDCASE 2026 Task 2 evaluation set, Zenodo, 2026. [Online]. Available: `https://zenodo.org/records/20485251`

[3] G. Jocher and J. Qiu, "Ultralytics YOLO11," 2024. [Online]. Available: `https://github.com/ultralytics/ultralytics`

[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. ICLR*, 2021.

[5] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF CVPR*, 2022.