GISP@HEU'S SUBMISSION FOR TASK 2: A CONSENSUS-ENSEMBLE YOLO SYSTEM FOR TEMPORAL DETECTION OF ANTARCTIC WHALE CALLS

Technical Report

Tong Ye[1], *Hongning Liu*[2], *Feiyang Xiao*[1], *Qiaoxi Zhu*[3], *Jian Guan*[1*]

[1]Group of Intelligent Signal Processing, Harbin Engineering University, Harbin, China
[2]School of Software, Dalian University of Technology, Dalian, China
[3]University of Technology Sydney, Ultimo, Australia

ABSTRACT

This report presents our submission to Task 2 of the BioDCASE 2026 Challenge, which addresses the temporal detection of whale calls in long underwater recordings with strongly labelled annotations. Rather than detecting events directly on the time axis, we recast the problem as 2D object detection on spectrogram images and reuse the YOLO11 detector, mapping predicted boxes back to time intervals. We submit four systems. System-1 is a single spectrogram-detection model, and System-2 is a consensus ensemble built on the System-1 backbone. System-3 augments the detection heads with a learnable edge-enhancement module, while System-4 replaces the temporal-difference channel with an instantaneous-frequency representation. On the BioDCASE 2026 Task 2 validation set, System-1 reaches a macro F1 of 0.672 and System-2 improves it to 0.679 under the official metric. System-3 and System-4 achieve macro F1 scores of 0.664 and 0.668 respectively.

***Index Terms*—** Bioacoustics, Sound Event Detection, Feature Enhancement, Ensemble Learning

1. INTRODUCTION

BioDCASE 2026 Task 2 targets the temporal detection of three whale call types in long passive-acoustic recordings collected at several deployment sites [1]. Passive acoustic monitoring (PAM) has become a standard tool for studying marine mammals, and deep learning is now widely used toanalyse such recordings [2]. The dataset contains seven annotated call types from two species, namely the Antarctic blue whale Z-Call, A-Call, B-Call, and D-Call, as well as the fin whale 20 Hz Pulse, its overtone variant, and the 40 Hz Downsweep. For evaluation, these seven types are merged into three classes: bmabz combines the blue-whale Z-Call, A-Call and B-Call; d corresponds to the D-Call; and bp groups the three fin-whale vocalisation types. For each file and class, predictions are matched to previously unmatched ground-truth intervals using 1D temporal IoU. A prediction is counted as correct if its maximum IoU with an unmatched ground-truth interval is strictly greater than 0.3; the matched ground-truth interval is then marked as detected and cannot be matched again. The resulting per-class TP, FP, and FN counts are used to compute precision and recall, from which we derive the class-wise and macro F1 scores.

The central design choice underlying our work is to avoid building a 1D temporal detector from scratch and instead treat detection as 2D object detection on the spectrogram. Treating audio events as patterns on a time-frequency image and processing them with deep networks is a well-established paradigm in sound event detection [3], which lets us reuse the strong and well-engineered YOLO family [4, 5]. Building on this idea, we submit four systems: a single spectrogram-detection model (System-1), a consensus ensemble of several such models (System-2) that further improves robustness across whale call types and deployment sites, an edge-enhanced variant (System-3) that injects learnable edge filters before the detection heads, and an instantaneous-frequency variant (System-4) that replaces the input temporal-difference channel with a phase-derived feature encoding frequency-modulation slopes.

*Corresponding author.

2. PROPOSED SYSTEM

2.1. System-1

System-1 is a single spectrogram-detection model. Each recording is converted to 224-bin log-mel spectrograms. The spectrogram is tiled, a YOLO11 detector predicts call boxes, and boxes are mapped back to time intervals and aggregated before per-class confidence thresholding. Since the largest whale call spans only about 144 pixels on the spectrogram, we remove the P5/32 head and add a P2/4 head, concentrating detection capacity on small and medium targets.

2.2. System-2

System-2 is a consensus ensemble built on top of System-1. It combines several detectors that share the same System-1 backbone, but they are trained with different strategies, including different random seeds, light data augmentation [6], and Gaussian-mixture feature-enhancement [7] fine-tuning specialized for the difficult d class. As a result, their predictions are complementary. The member predictions are merged using a weighted interval fusion method adapted from weighted box fusion [8] for the 1D temporal detection setting. The key idea is consensus voting: intervals supported by multiple independent members are assigned higher confidence, whereas isolated detections are down-weighted. This improves precision while preserving the diversity of the ensemble members. The fusion is applied on a per-class basis only when it brings performance gains, and the fused predictions are then filtered with class-specific confidence thresholds to produce the final submission.

2.3. System-3

System-3 shares the same backbone and training recipe as System-1 but inserts a learnable edge-enhancement layer at each detection branch (P2, P3, P4) immediately before the Detect head, aiming to sharpen the frequency-band boundaries of whale calls on the spectrogram.

2.4. System-4

System-4 keeps the model architecture identical to System-1 but replaces the temporal-difference B channel in the input spectrogram with an instantaneous-frequency (IF) proxy derived from the STFT phase, which directly encodes frequency-modulation slopes of whale calls while preserving the R and G channels.

Table 1: Per-class and macro F1 on the validation set for the four submitted systems.

Class	Baseline	System-1	System-2	System-3	System-4
`bmabz`	0.369	0.758	0.751	0.754	**0.762**
`d`	0.081	0.635	**0.656**	0.615	0.630
`bp`	0.198	0.624	**0.630**	0.623	0.611
macro	0.360	0.672	**0.679**	0.664	0.668

3. EXPERIMENTS AND RESULTS

3.1. Dataset

The BioDCASE 2026 Task 2 development set [1] is derived from the ATBFL library [9], one of the largest annotated datasets in marine bioacoustics, comprising underwater recordings collected around Antarctica between 2005 and 2017. It contains 6591 audio files (6004 train, 587 validation) totalling approximately 1880 hours from 11 site-year deployments. Seven call types from Antarctic blue whales and fin whales are annotated; for evaluation they are merged into three classes. The setup is multi-label as calls may overlap in time. We report results on the official validation split.

3.2. Results

Each recording is converted into a sequence of overlapping spectrogram images, on which the YOLO detectors operate, and the predicted boxes are then mapped back to time intervals. The detectors are trained on the training split and kept frozen during inference. For System-2, their predictions are fused as described in Section 2.2. System-3 uses the same inference pipeline but with edge-enhancement layers active in the detection heads, and System-4 operates on spectrograms whose B channel is replaced by the instantaneous-frequency proxy. All systems then apply per-class confidence thresholds. Performance is evaluated using the official metric, where a prediction is counted as correct if its temporal IoU with a ground-truth interval exceeds 0.3, and the macro F1 over the three classes is reported.

Table 1 reports the per-class and macro F1 scores of all four systems on the validation set. System-2 achieves the highest macro F1 of 0.679, showing that the consensus ensemble provides further gains over the already strong single System-1 model. System-3 adopts an edge-enhancement strategy, while System-4 uses an instantaneous-frequency input representation, and they achieve macro F1 scores of 0.664 and 0.668, respectively. Although their standalone performance is lower than that of System-1, they explore two complementary directions, namely architectural enhancement and input-representation design.

4. CONCLUSION

We presented four systems for BioDCASE 2026 Task 2 that frame temporal whale call detection as spectrogram object detection with YOLO11. System-1 is a single spectrogram-detection model reaching a validation macro F1 of 0.672, and System-2 is a consensus ensemble that improves it to 0.679. System-3 introduces learnable edge-enhancement layers before the detection heads and achieves a macro F1 of 0.664, while System-4 replaces the temporal-difference input channel with an instantaneous-frequency representation and reaches 0.668. The remaining headroom lies mainly in the recall of the d and bp classes and their cross-site generalization, which we leave for future work.

5. REFERENCES

[1] L. Jean-Labadye, C. Parcerisas, B. Miller, P. Carvaillo, G. Dubus, A. Gros-Martial, A. Marmoret, I. Moummad, P. Nguyen Hong Duc, P.-Y. Raumer, E. Schall, and D. Cazau, "2026 BioDCASE Task 2 : Development set," Mar. 2026.

[2] D. Stowell, "Computational bioacoustics with deep learning: a review and roadmap," *PeerJ*, vol. 10, 2021.

[3] E. Cakir, G. Parascandolo, T. Heittola, H. Huttunen, T. Virtanen, E. Cakir, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 25, no. 6, p. 1291–1303, June 2017.

[4] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.

[5] R. Khanam and M. Hussain, "YOLOv11: An overview of the key architectural enhancements," 2024.

[6] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020.

[7] H. Liu, P. Feng, M. Xie, D. Xu, J. Guan, G. He, and R. Zhang, "FPN with GMM based feature enhancement strategy for object detection in remote sensing images," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2024, pp. 3420–3424.

[8] R. Solovyev, W. Wang, and T. Gabruseva, "Weighted boxes fusion: Ensembling boxes from different object detection models," *Image and Vision Computing*, vol. 107, p. 104117, 2021.

[9] B. S. Miller, K. M. Stafford, I. Van Opzeeland, D. Harris, F. Samaran, A. Širovi'c, S. Buchan, K. Findlay, N. Balcazar, S. Nieukirk, E. C. Leroy, M. Aulich, F. W. Shabangu, R. P. Dziak, W. Lee, and J. Hong, "An Annotated Library of Underwater Acoustic Recordings for Testing and Training Automated Algorithms for Detecting Antarctic Blue and Fin Whale Sounds," 2020.