# JOINT WHALE CALL DETECTION AND RECORDING LOCATION CLASSIFICATION USING MULTITASK LEARNING FOR BIODCASE CHALLNEGE 2026 TASK 2

Technical Report

*Daito Isoda*[1], *Ryoko Arita*[1], *Yuki Okamoto*[1*],
*Noriyuki Tonami*[2], *Yuki Saito*[1], *Hiroshi Saruwatari*[1],

[1] The University of Tokyo, Japan,
[2] NEC Corporation, Japan,
y-oka@g.ecc.u-tokyo.ac.jp

## ABSTRACT

Passive Acoustic Monitoring (PAM) is widely used to monitor whale populations, but manually detecting whale calls from large-scale recordings is labor-intensive and difficult to scale. In this paper, we propose a multitask learning framework that jointly performs whale call detection and recording location classification, exploiting the geographical imbalance observed in whale call distributions across recording locations. The proposed model is based on YOLO and consists of a shared feature extraction network, a whale call detection branch, and an auxiliary recording location classification branch. We evaluate the proposed method on the BioDCASE Challenge 2026 Task 2 dataset and compare it with a baseline model trained without the auxiliary task. Experimental results show that the proposed method improves the F1-score by up to 0.06 over the baseline model. These findings suggest that incorporating location-dependent information through multitask learning can be beneficial for whale call detection.

***Index Terms***— Whale call detection, recording location classification, multitask learning

## 1. INTRODUCTION

Whale populations have declined dramatically, primarily due to overexploitation through whaling. As a result, several species, including the blue whale (Balaenoptera musculus), are currently classified as endangered. To monitor whale populations, the recording of whale calls using Passive Acoustic Monitoring (PAM) has been widely employed [1]. PAM enables the long-term observation of whales by continuously recording underwater acoustic signals. However, manually detecting whale calls from the recorded acoustic data is labor-intensive and difficult to scale. To address this limitation, considerable research has focused on the automatic detection of whale calls using deep learning techniques [2, 3, 4].

Several publicly available datasets have been released for whale call detection. Besides annotations of whale call types and their occurrence intervals, these datasets commonly provide environmental metadata, including recording location and depth. The BioDCASE Challenge 2026 Task 2 dataset [1] contains seven whale call classes, including Antarctic blue whale calls (BmA, BmB, BmD, and BmZ) and fin whale calls (Bp20, Bp20plus, and BpD). Using the location metadata of the this dataset, we analyzed the occurrence frequency of whale calls across recording locations. As shown in Figure 1, the distribution of whale calls exhibits a noticeable imbalance among different geographical areas. This observation suggests that recording location information may serve as a useful contextual cue for whale call detection, motivating the incorporation of location-related information into the detection framework.

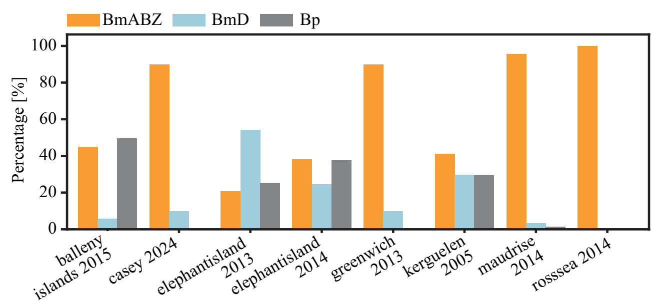


Figure 1: Distribution of whale calls for each recording location. BmABZ, BmD, and Bp denote Antarctic blue whale A-, B-, and Z-calls, Antarctic blue whale D-calls, and fin whale calls, respectively.

In this paper, we propose a multitask learning framework that jointly performs whale call detection and recording location classification to exploit geographical variations in whale call distributions. By learning these two tasks simultaneously, the model is expected to leverage location-dependent patterns in whale call occurrences, thereby improving the performance of whale call detection. Specifically, the auxiliary task of recording location classification provides contextual information regarding the likelihood of observing particular whale calls in different regions, which may assist the primary detection task. Experimental results indicate that the proposed approach can improve whale call detection performance.

## 2. PROPOSED METHOD

The overall architecture of the proposed method is shown in Figure 2. The proposed model extends the YOLO v11 [5]-based whale call detection baseline used in the BioDCASE Challenge 2026[1] by

*This work was supported by Support Center for Advanced Telecommunications Technology.

[1] `https://github.com/marinebioCASE/task2_2025_2026/tree/main/baselines`

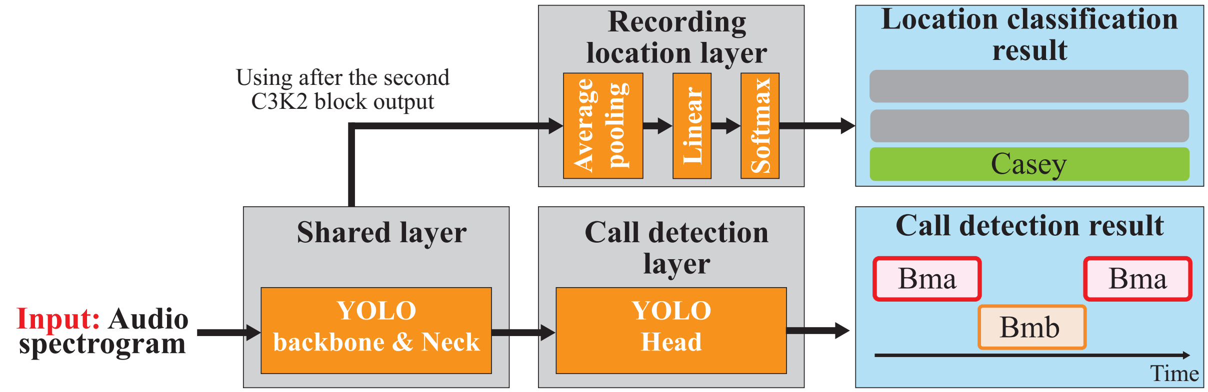


Figure 2: Overview of model architecture

Table 1: Detection results for validation set when $\tau$ is set to 0.1

| Method | Recall | Precision | F1-Score | Location accuracy |
|---|---|---|---|---|
| Baseline | 0.494 | **0.537** | 0.515 | - |
| Proposed | | | | |
| $\lambda = 0.01$ | 0.711 | 0.410 | **0.520** | **31.8%** |
| $\lambda = 0.1$ | **0.744** | 0.326 | 0.454 | 1.58% |
| $\lambda = 1.0$ | 0.730 | 0.343 | 0.467 | 26.9% |
| $\lambda = 10.0$ | 0.654 | 0.403 | 0.499 | 5.1% |

Table 2: Detection results under different threshold settings. In the proposed method, $\lambda$ is set to 0.01

| Method | Recall | Precision | F1-Score |
|---|---|---|---|
| Baseline | | | |
| $\tau = 0.1$ | 0.494 | 0.537 | 0.515 |
| $\tau = 0.2$ | 0.354 | 0.726 | 0.476 |
| $\tau = 0.3$ | 0.216 | **0.835** | 0.343 |
| Proposed | | | |
| $\tau = 0.1$ | **0.711** | 0.410 | 0.520 |
| $\tau = 0.2$ | 0.581 | 0.571 | **0.576** |
| $\tau = 0.3$ | 0.447 | 0.709 | 0.548 |

introducing an auxiliary recording location classification branch. The resulting architecture consists of a shared feature extraction network, a whale call detection branch, and a recording location classification branch.

Given an input audio spectrogram, the shared layer first extracts acoustic representations using the YOLO backbone and neck. The extracted features are then fed into two task-specific branches. The primary branch is the call detection layer, which corresponds to the original YOLO detection head and predicts the temporal positions and categories of whale calls. This branch is optimized using the whale call detection objective.

To exploit geographical differences in whale call distributions, we introduce an auxiliary recording location classification branch. Specifically, the output feature map of the second cross stage partial with kernel size 2 (C3K2) block in the backbone is used as the input to the location classification layer. Global average pooling is first applied to aggregate the feature map into a fixed-length feature vector, followed by a fully connected layer and a softmax classifier to predict the recording location of the input audio.

The model is trained in a multitask learning manner by jointly optimizing the whale call detection task and the recording location classification task. Through the auxiliary supervision of location classification, the shared feature extractor is encouraged to learn location-dependent acoustic characteristics and occurrence patterns of whale calls. As a result, the learned location information is expected to improve the performance of whale call detection.

The overall loss function is defined as

$$L = L_{\mathrm{det}} + \lambda L_{\mathrm{loc}}, \tag{1}$$

where $L_{\mathrm{det}}$ and $L_{\mathrm{loc}}$ denote the whale call detection loss and the recording location classification loss, respectively, and $\lambda$ is a weighting coefficient that controls the contribution of the auxiliary task. In this study, $L_{\mathrm{loc}}$ is implemented using the cross-entropy loss. In this study, $L_{\mathrm{loc}}$ is implemented using the cross-entropy loss. For $L_{\mathrm{det}}$, we employ the same detection loss function as that used in the BioDCASE Challenge 2026 Task 2 baseline model.

## 3. EXPERIMENTS

### 3.1. Experimental setup

We evaluated the proposed method on the BioDCASE Challenge 2026 Task 2 dataset, which contains underwater acoustic recordings collected from eight recording locations: Balleny Islands (2015), Casey (2024), Elephant Island (2013, 2014), Greenwich (2013), Kerguelen (2005), Maud Rise (2014), and Ross Sea (2014). The dataset includes annotations for three whale call types: BmABZ, BmD, and Bp. Each audio recording was converted into a spectrogram with a sampling rate of 250 Hz, using a window size of 512 and a hop size of 20.

The model was trained for 20 epochs with a batch size of 32, using the SGD optimizer with a learning rate of 0.001. The weighting coefficient $\lambda$ in Eq. (1) was varied among 0.01, 0.1, 1.0, and 10.0 to investigate its effect on detection performance. As a baseline, we compared it to the YOLO-based model from BioDCASE Challenge 2026 Task 2.

For evaluation, predicted whale calls were matched with ground-truth annotations based on an Intersection over Union (IoU) threshold of 0.3. Recall, precision, and F1-score were computed using a detection confidence threshold $\tau$. To examine the effect of this threshold, we evaluated both the baseline and proposed methods at $\tau = 0.1, 0.2$, and $0.3$. Location classification accuracy was computed as the proportion of correctly predicted recording locations among all validation samples.

### 3.2. Experimental results

Table 1 shows the detection results on the validation set for the baseline and the proposed method with different values of the weighting coefficient $\lambda$. Compared to the baseline (F1-score $= 0.515$), the proposed method with $\lambda = 0.01$ achieved a slightly higher F1-score ($0.520$), with a substantial improvement in recall (from $0.494$ to $0.711$) at the cost of a lower precision (from $0.537$ to $0.410$). For larger values of $\lambda$ ($0.1$ and $1.0$), the F1-score decreased to $0.454$ and $0.467$, respectively, despite the recall remaining higher than that of the baseline. With $\lambda = 10.0$, the F1-score ($0.499$) was closer to but still below that of the baseline.

Regarding the auxiliary task, the location classification accuracy varied substantially depending on $\lambda$, ranging from $1.58\%$ ($\lambda = 0.1$) to $31.8\%$ ($\lambda = 0.01$). Notably, the highest detection F1-score ($\lambda = 0.01$) corresponded to the highest location classification accuracy, suggesting a possible relationship between the model's ability to learn location-dependent representations and its detection performance. However, this relationship was not consistently observed across all values of $\lambda$, as the location accuracy did not increase monotonically with detection performance.

Table 2 shows the detection results of the baseline and the proposed method ($\lambda = 0.01$) under different detection confidence thresholds $\tau$. For both methods, increasing $\tau$ led to a decrease in recall and an increase in precision, reflecting the expected trade-off between these two metrics. The proposed method consistently outperformed the baseline in terms of F1-score across all threshold settings, with the largest improvement observed at $\tau = 0.2$ ($0.576$ vs. $0.476$) and $\tau = 0.3$ ($0.548$ vs. $0.343$). These results indicate that the proposed method maintains a more favorable recall-precision balance than the baseline, particularly at higher threshold settings where the baseline's recall drops sharply.

Overall, these results suggest that incorporating the auxiliary recording location classification task can improve whale call detection performance, particularly in terms of recall and the overall recall-precision balance, although the optimal value of $\lambda$ requires careful tuning.

## 4. CONCLUSION

In this paper, we proposed a multitask learning framework for whale call detection that jointly learns whale call detection and recording location classification, motivated by the geographical imbalance observed in whale call distributions across recording locations. Experimental results demonstrated that the proposed method consistently improved the F1-score over the baseline across different detection confidence thresholds, mainly owing to a substantial improvement in recall. In particular, with $\lambda = 0.01$, the proposed method achieved the highest location classification accuracy among the tested settings, suggesting a possible link between learning location-dependent representations and improved detection performance. However, this relationship was not consistently observed for other values of $\lambda$, indicating that the contribution of the auxiliary task is sensitive to its weighting coefficient. Future work includes investigating additional auxiliary tasks that exploit other forms of environmental metadata to further improve whale call detection performance.

## 5. REFERENCES


[1] B. S. Miller, N. Balcazar, S. Nieukirk, E. C. Leroy, M. Aulich, F. W. Shabangu, R. P. Dziak, W. S. Lee, and J. K. Hong, "An open access dataset for developing automated detectors of antarctic baleen whale sounds and performance evaluation of two commonly used detectors," *Scientific Reports*, vol. 11, no. 1, p. 806, 2021.

[2] E. Schall, I. I. Kaya, E. Debusschere, P. Devos, and C. Parcerisas, "Deep learning in marine bioacoustics: a benchmark for baleen whale detection," *Remote Sensing in Ecology and Conservation*, vol. 10, no. 5, pp. 642–654, 2024.

[3] M. J. Kim, J. Lee, Y. Cho, W.-K. Kim, J. Park, D. Lee, and H. S. Bae, "Automated detection and classification of marine species vocalizations using a yolo-based deep learning framework," *Ecology and Evolution*, vol. 16, no. 4, p. e73466, 2026.

[4] R. Zhang, Y. Shi, L. Yang, and Z. Zhou, "Classification of whale call signals based on the improved yolov5 model," in *Proc. 6th International Conference on Information Communication and Signal Processing (ICICSP)*, 2023, pp. 429–433.

[5] R. Khanam and M. Hussain, "Yolov11: An overview of the key architectural enhancements," *arXiv preprint:2410.17725*.