Supervised vs Self-Supervised Representation Learning for Fin Whale Vocalization Detection

Anonymous Author(s)
Affiliation
Address
email

Abstract

Fin whales produce low-frequency vocalizations critical for monitoring but are often masked by anthropogenic noise. While supervised detectors perform well, they require costly labels and degrade under noise or data scarcity. We present the first application of self-supervised learning (SSL) to fin-whale detection, combining contrastive predictive coding with an amplitude-aware encoder. Across datasets collected in the Norwegian Sea and the Mediterranean Sea, SSL models outperform supervised Transformers in low-label and low SNR regimes and transfer effectively across regions. Embedding visualizations further show robust class separability. These results highlight SSL as a scalable approach for passive acoustic monitoring, reducing annotation needs and paving the way for scalable, label-efficient acoustic monitoring across diverse marine habitats.

1 Introduction

Fin whales (*Balaenoptera physalus*) produce stereotyped low-frequency vocalizations, near 20 Hz and 125 Hz, that propagate over hundreds of kilometers, enabling long-range pelagic communication Croll et al. [2002], Tyack [2008], Best et al. [2022]. Since only males sing, vocalizations are closely related to reproduction, making passive acoustic monitoring (PAM) a key tool for population assessment and mitigation Croll et al. [2002]. However, fin whale vocalizations overlap with intense anthropogenic noise (e.g., shipping), leading to masking and reduced communication range Duarte et al. [2021], Castellote et al. [2012].

Supervised deep-learning methods have recently shown strong performance in fin-whale detection, but they rely on costly expert labels and degrade under label noise, low SNR, or limited data. Self-supervised learning (SSL) offers a promising alternative. By learning robust embeddings from unlabeled audio, it reduces annotation needs and improves detection and transferability. Advances in contrastive and predictive audio SSL Oord et al. [2018], Baevski et al. [2020], Stowell [2022] highlight its potential for bioacoustics.

We compare a supervised lightweight transformer encoder and SSL-based detectors on two datasets, evaluate seeds with uncertainty, and analyze learned embeddings with t-SNE. We further test robustness under noise, data scarcity, and cross-site transfer between two noisy regions: the Norwegian Sea and the Mediterranean Sea.

2 Related Work

2.1 Supervised Detection of Fin Whale Songs

Supervised deep learning has already shown strong results on fin whale vocalizations. Best et al. [2022] introduced an active-learning framework with lightweight CNNs to detect stereotyped 20

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

Hz calls in the Northwestern Mediterranean Sea. Their model, with only 36k trainable parameters, achieved outstanding performance AUROC scores (0.992 at Bombyx, 0.997 at Boussole) on the FinWhaleSong dataset, while remaining efficient for deployment in resource-limited monitoring settings. These results demonstrate the potential of supervised CNNs for detection in noisy soundscapes; however, their dependence on expert validation limits scalability to large, unlabeled datasets.

2.2 Self-Supervised Learning for Bioacoustics

Supervised methods depend on expensive and noisy labels, while self-supervised learning (SSL) leverages unlabeled audio to learn robust representations. SSL frameworks such as CPC Oord et al. [2018] and wav2vec 2.0 Baevski et al. [2020] achieve near-supervised performance in speech, and in bioacoustics they improve detection, support few-shot classification, and transfer effectively from human to animal vocalizations Stowell [2022], Moummad et al. [2024], Sarkar and Doss [2023, 2025]. However, SSL has not yet been applied to fin-whale calls. To fill this gap, we release an annotated dataset and compare supervised and SSL-based detectors.

3 Methodology

3.1 Supervised Transformer Encoder Model

First, we transform the input signal into a low-dimensional spectrogram using an STFT.

Our supervised baseline is a lightweight Transformer encoder applied directly to spectrogram inputs. Each column $S(:,t) \in \mathbb{R}^F$, corresponding to F frequency bins at time step t, is treated as the input embedding at that step. This avoids the need for an explicit embedding layer, following recent work in audio Transformers Gong et al. [2021], Dong et al. [2018].

The encoder consists of 4 Transformer blocks, each with 4 heads of self-attention and a feed-forward layer, combined with residual connections and layer normalization. Unlike most Transformer models, we omit explicit positional encodings, allowing the model to rely purely on spectral–temporal patterns and the receptive field of the attention heads to capture dependencies Su et al. [2024].

The output sequence is projected through a linear layer to obtain frame-level probabilities $\hat{y}_t$, which are then flattened and aggregated into a binary classification of presence/absence.

3.2 Self-Supervised Method

For the self-supervised model, we adopt the contrastive predictive coding framework Oord et al. [2018], which combines a Sinc-based Ravanelli and Bengio [2018] front-end, a convolutional encoder with amplitude-aware normalization, and a uni-directional gated recurrent unit (GRU) Dey and Salem [2017] context model. The model takes raw waveform windows and learns to predict future latent representations K steps ahead using the InfoNCE objective.

Given an input waveform $x_t \in \mathcal{X}$, the encoder is defined as a mapping $g_{enc} : \mathcal{X} \to \mathcal{Z}$ parameterized by a five-layer convolutional network, producing representations $z_t = g_{enc}(x_t)$. The first convolutional layer is a **Sinc-based filter** constrained to learn band-pass filters $g[n, f_l, f_h] = 2f_h \cdot \mathrm{sinc}(2\pi f_h n) - 2f_l \cdot \mathrm{sinc}(2\pi f_l n)$, where $\mathrm{sinc}(x) = \frac{\sin(x)}{x}$. Each filter learns only the low and high cutoff frequencies (f_l, f_h), reducing parameters while yielding physically interpretable filters. For fin-whale vocalizations, this design ensures the encoder emphasizes ecologically relevant bands and suppresses redundant signals. An ablation of the SincNet is provided in Appendix F.2.

We propose **Batch-RMS Normalization (BRN)** to better preserve amplitude information in fin-whale detection. While Layer and Group Normalization Ba et al. [2016], Wu and He [2018] used in SSL frameworks Baevski et al. [2020], Hsu et al. [2021], Chen et al. [2022] stabilize training, they remove mean and scale, enforcing amplitude invariance which is detrimental when amplitude is an ecologically meaningful cue. Batch Normalization (BN) Ioffe and Szegedy [2015] retains such cues but is unstable with small batches and under distribution shift Li et al. [2019], Yang et al. [2022]. BRN interpolates between BN and RMSNorm Zhang and Sennrich [2019] with a learnable gate ρ: $\mathrm{BRN}(x) = \big(\rho \cdot \mathrm{BN}(x) + (1-\rho) \cdot \mathrm{RMSNorm}(x)\big) \odot \gamma + \beta$, where γ, β are trainable affine terms. BRN thus preserves amplitude-sensitive cues essential for pulse detection, while inheriting the stability and robustness of modern normalization. Ablation results are provided in Appendix F.1.

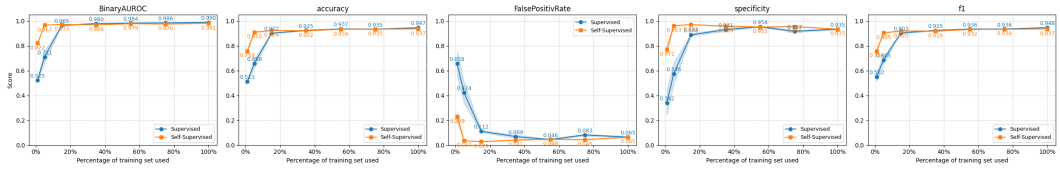

Figure 1: Performance of supervised method and SSL method under varying training set sizes on Seglvik Fjords dataset, demonstrating the robustness of our SSL compared to Supervised.

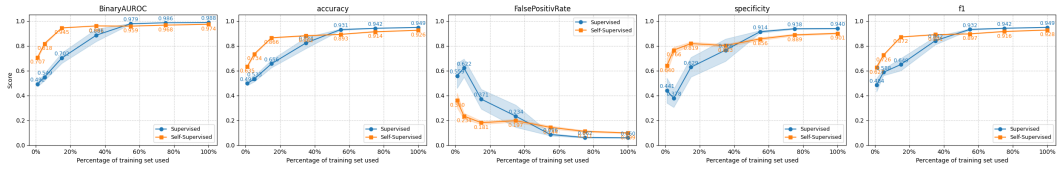

Figure 2: Performance of supervised method and SSL method under varying training set sizes on Mediterranean Sea.

With amplitude-sensitive features extracted by the encoder, the model employs an autoregressive context module to capture temporal dependencies and optimize predictive representations. Specifically, we use a GRU $g_{ar} : \mathcal{Z} \to \mathcal{C}$, which summarizes past representations $z_{\leq t}$ into a context vector $c_t = g_{ar}(z_{\leq t})$. To predict the future, K linear predictors $\{W_k\}_{k=1}^{K}$ generate $\hat{z}_{t+k} = W_k c_t$. Training follows the InfoNCE objective Oord et al. [2018], where the model learns to identify the true z_{t+k} among negatives $\{z_j\}_{j=1}^{N}$.

$$\mathcal{L}_k = -\log \frac{\exp\left(\text{sim}(\hat{z}_{t+k}, z_{t+k})/\tau\right)}{\sum_{j \in \{t+k\} \cup \mathcal{N}} \exp\left(\text{sim}(\hat{z}_{t+k}, z_j)/\tau\right)}, \tag{1}$$

with $\text{sim}(u, v) = u^\top v$ and temperature τ. The final loss is averaged across steps: $\mathcal{L} = \frac{1}{K} \sum_{k=1}^{K} \mathcal{L}_k$, ensuring the model jointly optimizes predictions over multiple future steps.

After pretraining, we freeze the backbone model, and use the GRU outputs c_t as input to a lightweight classifier consisting of a single convolutional layer and a linear head, trained in a supervised manner on labeled fin-whale pulse data. Detailed pipeline of the SSL model is shown in Appendix G.

3.3 Datasets

We evaluate on two fin-whale datasets. **Seglvik Fjords (Norway)** comprises over 1500 hours of recordings collected in a noisy Arctic environment with seasonal whale aggregations. Training labels were obtained through YOLOv5 Jocher et al. [2022], while validation and test sets were fully manually curated. Each sample is represented as a 3-s segment centered on the 125 Hz pulse band. **Mediterranean Sea (Bombyx and Boussole)** uses the FinWhaleSong dataset Best et al. [2022], containing nearly 3700 annotated 20 Hz pulses generated by their CNN. We create a train/val/test split as described in Appendix A and extract 5-s segments with one simple temporal augmentation. Further details on data collection, annotation, and preprocessing are provided in Appendix A.

4 Experiments

4.1 Experimental Setup & Protocol

Our downstream task is binary pulse detection: given an audio segment, the model predicts fin-whale presence or absence. We evaluate with **Accuracy**, **AUROC**, **F1**, **Recall**, **Specificity**, and **FPR**, averaging results over 10 random seeds for reproducibility. Models are tested on both **Seglvik Fjords** and **Mediterranean Sea** datasets using full training sets and subsets ranging from 1–100% of labeled data. To further probe noise robustness, we conduct an auxiliary Seglvik experiment (Appendix D) where training subsets are restricted to low-SNR samples ($\tau \in -4, 0, 2, 4$ dB) while evaluation

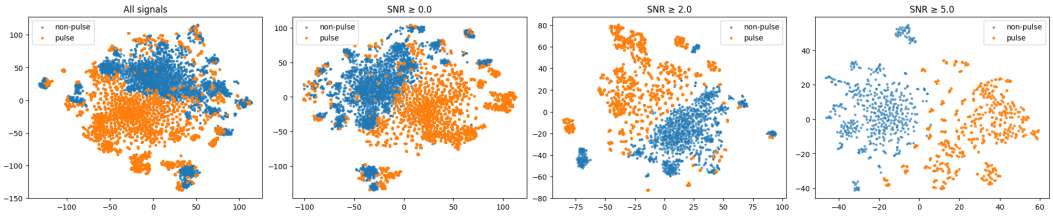

Figure 3: **t-SNE embeddings of the SSL model at different SNR thresholds with the Seglvik dataset.** Left to right: all signals, SNR ≥ 0, ≥ 2, and ≥ 5 dB.

remains on the full testset. Additional implementation details, including optimizer settings and model sizes, are provided in Appendix B.

4.2 Result & Analysis

On the **Seglvik Fjords** dataset, the supervised Transformer Encoder and SSL-pretrained encoder perform comparably with full supervision (Table 1, 2). In low-label regimes, however, SSL yields consistently higher accuracy, F1, and specificity, while nearly halving false positives. As shown in Fig. 1, with only 1% of labeled data, the SSL model already achieves strong performance, indicating that pretraining provides robust, discriminative features for detection under annotation scarcity. As labeled data increase, the supervised model closes the gap, slightly surpassing the frozen SSL encoder, which is expected given the absence of fine-tuning. These results demonstrate the potential of SSL to substantially reduce labeling requirements while suggesting that further fine-tuning could unlock additional gains in high-label settings.

To assess transferability and generalization, we used the Seglvik-pretrained model as the feature extractor and trained a lightweight classifier on the **Mediterranean Sea** dataset, compared to a supervised Transformer Encoder trained from scratch. As shown in Fig. 2, SSL substantially outperforms supervised model with 1–15% of labeled data and achieves strong performance, demonstrating that it has captured generalizable features of fin-whale vocalizations. With more annotations, the supervised model overtakes the frozen SSL encoder as freezing limits the SSL model's ability to adapt. Importantly, this result addresses a core challenge in bioacoustics: acoustic conditions differ drastically across regions, making large annotated datasets impractical for every site. SSL thus offers a scalable solution by reducing reliance on expert labels and enabling effective cross-site monitoring.

After comparing the SSL model and the supervised model, we next investigate the separability of embeddings from the frozen pretrained encoder. We input 3-s audio segments from the Seglvik Fjords testset into the frozen pretrained model, extract embeddings with RMS-weighted pooling, and visualize them with t-SNE. As shown in Fig. 3, pulse and non-pulse samples are already clearly separated in 2D space, indicating that the model encodes discriminative features for detection without task-specific fine-tuning. When stratifying the testset by SNR, the boundaries between classes become progressively sharper, demonstrating robustness to noise and enhanced separability with higher SNR. For the supervised model, which does not yield explicit embeddings, we instead use the output of the last transformer encoder block to visualize contextual representations of pulses and non-pulse as described in Appendix E.1.

5 Discussion & Conclusion

We introduced the first study applying self-supervised learning to fin-whale vocalization detection, comparing a lightweight supervised Transformer Encoder with a CPC-based encoder augmented by SincNet and amplitude-aware normalization. Across two large-scale datasets, our results show that SSL substantially improves performance in low-label regimes and provides transferable representations across geographically distinct acoustic environments. These findings highlight SSL as a practical and scalable approach for passive acoustic monitoring of whales, alleviating the reliance on costly expert labels and enabling more efficient deployment across diverse sites. Future work could extend this direction by incorporating fine-tuning, multi-species training, and large-scale pretraining to develop general-purpose bioacoustic representation models.

References

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

Paul Best, Ricard Marxer, Sébastien Paris, and Hervé Glotin. Temporal evolution of the mediterranean fin whale song. *Scientific reports*, 12(1):13565, 2022.

Manuel Castellote, Christopher W Clark, and Marc O Lammers. Acoustic and behavioural changes by fin whales (balaenoptera physalus) in response to shipping and airgun noise. *Biological Conservation*, 147(1):115–122, 2012.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6): 1505–1518, 2022.

Donald A Croll, Christopher W Clark, Alejandro Acevedo, Bernie Tershy, Sergio Flores, Jason Gedamke, and Jorge Urban. Only male fin whales sing loud songs. *Nature*, 417(6891):809–809, 2002.

Rahul Dey and Fathi M Salem. Gate-variants of gated recurrent unit (gru) neural networks. In *2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS)*, pages 1597–1600. IEEE, 2017.

Linhao Dong, Shuang Xu, and Bo Xu. Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5884–5888. IEEE, 2018.

Carlos M Duarte, Lucille Chapuis, Shaun P Collin, Daniel P Costa, Reny P Devassy, Victor M Eguiluz, Christine Erbe, Timothy AC Gordon, Benjamin S Halpern, Harry R Harding, et al. The soundscape of the anthropocene ocean. *Science*, 371(6529):eaba4658, 2021.

Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*, 2021.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460, 2021.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015.

Glenn Jocher, Ayush Chaurasia, Alex Stoken, Jirka Borovec, Yonghye Kwon, Kalen Michael, Jiacong Fang, Zeng Yifu, Colin Wong, Diego Montes, et al. ultralytics/yolov5: v7. 0-yolov5 sota realtime instance segmentation. *Zenodo*, 2022.

Jialin Li, Xueyi Li, and David He. Domain adaptation remaining useful life prediction method based on adabn-dcnn. In *2019 Prognostics and System Health Management Conference (PHM-Qingdao)*, pages 1–6. IEEE, 2019.

Ilyass Moummad, Romain Serizel, and Nicolas Farrugia. Self-supervised learning for few-shot bird sound classification, 2024. URL `https://arxiv.org/abs/2312.15824`.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Mirco Ravanelli and Yoshua Bengio. Speaker recognition from raw waveform with sincnet. In *2018 IEEE spoken language technology workshop (SLT)*, pages 1021–1028. IEEE, 2018.

Eklavya Sarkar and Mathew Magimai. Doss. Can self-supervised neural representations pre-trained on human speech distinguish animal callers?, 2023. URL https://arxiv.org/abs/2305.14035.

Eklavya Sarkar and Mathew Magimai. Doss. Comparing self-supervised learning models pre-trained on human speech and animal vocalizations for bioacoustics processing, 2025. URL https://arxiv.org/abs/2501.05987.

Dan Stowell. Computational bioacoustics with deep learning: a review and roadmap. *PeerJ*, 10: e13152, 2022.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Peter L Tyack. Implications for marine mammals of large-scale changes in the marine acoustic environment. *Journal of Mammalogy*, 89(3):549–558, 2008.

Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

Tao Yang, Shenglong Zhou, Yuwang Wang, Yan Lu, and Nanning Zheng. Test-time batch normalization. *arXiv preprint arXiv:2205.10210*, 2022.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in neural information processing systems*, 32, 2019.

Appendix

A Datasets & Code

Seglvik Fjords (Norway). The Seglvik Fjords dataset was collected in northern Norway and comprises 1555 hours of recordings between November 12, 2022 and January 23, 2023. The site hosts seasonal aggregations of fin whales in an environment characterized by high variability in ambient noise due to shipping and other cetaceans. Recordings were made at a 190 kHz sampling rate, but since fin whale vocalizations occur at very low frequencies, all audio files were downsampled to 3.2 kHz. For this dataset, fin whale pulses are observed at $\sim$125 Hz.

Annotation was conducted in several stages. First, we manually labeled 207 fin whale pulses and 22 non-pulses. A YOLOv5 detector was then trained and applied to the entire 1555 hours of recordings, producing 160,161 candidate detections. To ensure a low label-noise, we retained only detections with a confidence score above 0.9, which reduced the training set from 160,161 to 4,129 pulses. Non-pulses are generated, in the same quantity as pulses, by avoiding pulse zones with all YOLO confidence levels. This constitutes the training dataset used in our experiments.

In addition to the YOLO-based training labels, we curated fully manual annotations for a more consistent test set and validation set. The test set contains 581 manually validated 125 Hz fin whale pulses, and the validation set contains 144 pulses. Each sample is extracted as a 3-s segment, not necessarily centered on the event and filtered with the frequency band 100–150 Hz. No data augmentation was applied for this dataset.

The full dataset can be downloaded from `https://drive.google.com/file/d/17WUR35hvPUp_wJpOSdN_DTxTEk39E4rY/view?usp=sharing`.

Mediterranean Sea (Bombyx and Boussole). We also evaluate on the FinWhaleSong dataset Best et al. [2022], which was generated using their CNN on the audio collected in the Northwestern Mediterranean: Bombyx and Boussole. Recordings were made with hydrophones and downsampled to 200 Hz, which is sufficient to capture the stereotyped $\sim$20 Hz pulses produced by fin whales. The initial dataset was manually annotated through an active learning cycle, yielding high-quality labels of pulse occurrences.

The published benchmark contains 39 hours of audio recordings and 3,688 annotations of fin whale pulses using their CNN. Since the public version is not pre-partitioned, we filtered by CNN confidence and randomly divided the dataset into training (1,031 pulses and 1,031 non-pulses), validation (103 pulses and 103 non-pulses), and test (97 pulses and 97 non-pulses) subsets. The generation of non-pulses follows the same approach as those generated with the Seglvik dataset.

For input representation, each sample is extracted as a 5-s segment non-centered. To enhance robustness, we applied simple data augmentation by shifting each segment once along the time axis. We also filtered the signal with the 15-35 Hz frequency band.

The full dataset can be downloaded from `https://drive.google.com/file/d/1OznWIDn0Ta_deoVzSiDqeCotQTUbmjJR/view?usp=sharing`.

Code. The code for preprocessing both datasets and training the model with evaluation can be downloaded from `https://drive.google.com/file/d/1ShGh0cGTkUHSBlX2PgHL-sCmGTzaHIBb/view?usp=drive_link`.

B Experimental Setup

Supervised training. For supervised models, we use the AdamW optimizer with weight decay $1e{-}5$, a OneCycleLR scheduler, and a binary cross-entropy loss. Training is performed for 20 epochs with a batch size of 32 on a single RTX 3090 GPU for 2 mins, followed, depending on the type of dataset, by an additional fine-tuning phase of four epochs on the validation set in order to adapt the decision limits or the calculation of the optimal threshold based on the results on the validation set. The models are deliberately compact, with 84k parameters on the Seglvik dataset and 60k on the Mediterranean Sea dataset with 4 encoder blocks and 4 heads, ensuring a fair comparison with prior CNN-based baselines Best et al. [2022].

Self-supervised pretraining & Downstream Task training. For SSL, we adopt Adam with an initial learning rate $1e{-}4$ and weight decay $1e{-}5$, together with a customized scheduler that dynamically adjusts the learning rate per step. Pretraining runs for 50 epochs on Seglvik's training dataset on $4\times$ RTX 3090 GPUs with a batch size of 512 for 3 days, yielding a backbone of 7M parameters. During downstream adaptation, the pretrained encoder and GRU are frozen, and only a lightweight classification head is trained. The downstream task training uses AdamW with cosine annealing with an initial learning rate $1e{-}3$, weight decay $1e{-}4$, and a batch size of 32 on a single RTX 3090 GPU for 2 mins.

C Evaluation Metrics for Supervised vs. SSL-pretrained (frozen) model.

Table 1 & Table 2 report metrics on the Seglvik dataset, trained with the full training set. Evaluation was performed using an automatically selected optimal threshold strategy based on the results of the validation set.

Model	Acc	AUROC	FPR	Spec	F1
Supervised	0.947 ± 0.001	0.990 ± 0.000	0.065 ± 0.006	0.935 ± 0.006	0.948 ± 0.001
SSL	0.937 ± 0.001	0.981 ± 0.001	0.065 ± 0.004	0.935 ± 0.004	0.937 ± 0.001

Table 1: **Seglvik Fjords results: supervised vs SSL.** Mean $\pm$ sem across 10 seeds.

Model	Recall	Precision	Loss
Supervised	0.959 ± 0.004	0.937 ± 0.005	0.372 ± 0.002
SSL	0.938 ± 0.003	0.936 ± 0.004	0.381 ± 0.002

Table 2: **Additional metrics on Seglvik Fjords dataset.** Mean $\pm$ sem across 10 seeds.

Table 3 & Table 4 present results on the Mediterranean dataset. Models were trained on the full training set, and evaluation was performed using an automatically selected optimal threshold strategy based on the results of the validation set.

Model	Acc	AUROC	FPR	Spec	F1
Supervised	0.951 ± 0.002	0.988 ± 0.001	0.059 ± 0.003	0.941 ± 0.003	0.951 ± 0.003
SSL	0.927 ± 0.002	0.974 ± 0.001	0.098 ± 0.005	0.902 ± 0.005	0.929 ± 0.002

Table 3: **Mediterranean results: supervised vs SSL.** Mean $\pm$ sem across 10 seeds.

Model	Recall	Precision	Loss
Supervised	0.960 ± 0.006	0.942 ± 0.002	0.369 ± 0.002
SSL	0.952 ± 0.003	0.907 ± 0.004	0.401 ± 0.002

Table 4: **Additional metrics on Mediterranean dataset.** Mean $\pm$ sem across 10 seeds.

D Transferability and Robustness Across SNR Conditions

To directly test model robustness under noisy training data, we construct training sets on the Seglvik dataset by applying a maximum SNR threshold, using only samples with SNR $\leq \tau$, where $\tau \in -4, 0, 2, 4$ dB. The test set remains unchanged and includes the full distribution of SNR levels. Each setting is repeated with ten random seeds for statistical reliability. As shown in Fig. 4, supervised performance is strongly impacted when trained on low-SNR subsets, with AUROC, accuracy, and F1 all suppressed, and false positive rates increasing sharply. In contrast, SSL maintains much higher performance across metrics, even when trained only on noisy data, and its advantage is

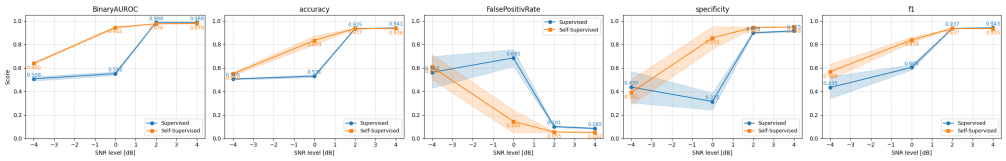

Figure 4: Performance of supervised and SSL-based models on Seglvik when training is restricted to noisy subsets with SNR $\leq \tau$, while testing is performed on the full test set. Supervised models degrade substantially as τ decreases, showing high false positive rates and reduced F1, whereas SSL remains comparatively stable and robust across thresholds. Results are averaged over 10 random seeds.

most pronounced at the lowest SNR thresholds. These results confirm that SSL can extract robust representations from noisy audio, consistent with our motivation that supervised models are vulnerable to label noise and low SNR conditions, whereas SSL leverages unlabeled structure to build more resilient features.

E Embedding visualization

E.1 t-SNE visualization for the supervised model

To analyze the internal representations of the supervised detector, we extract hidden states from the last transformer encoder block and project them into 2D space with t-SNE. As with the t-SNE visualization for the SSL model, we enrich the testset with five temporal-shift augmentations around the pulse to provide more stable visualizations and smoother clusters.

Figure 5 shows that pulse and non-pulse samples remain broadly separable, though the clusters are less geometrically distinct than for the frozen encoder. At low SNR, the overlap between classes is more pronounced, consistent with the higher classification difficulty. As SNR increases, class separation becomes clearer, and the two groups form compact clusters, mirroring the model's improved detection performance at higher signal quality.

This comparison highlights a key difference: pretrained embeddings encode discriminative features in a form that is directly separable even without supervision, while the supervised model organizes its internal states around the classification task, yielding more diffuse but still discriminative representations.

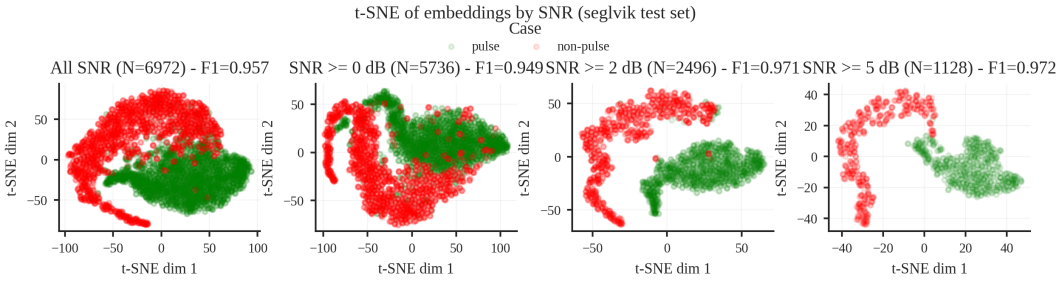

Figure 5: **t-SNE embeddings of the Supervised model at different SNR thresholds with the Seglvik dataset.** Left to right: all signals, SNR ≥ 0, ≥ 2, and ≥ 5 dB.

E.2 Embedding-Space Separability and Representative Signals

To provide qualitative examples of the data, we visualize embeddings from the frozen SSL encoder using t-SNE together with a few randomly selected audio segments from the Seglvik test set. As shown in Fig. 6, pulses and non-pulses form distinct regions in the embedding space, and the spectrograms illustrate representative signal patterns for each class.

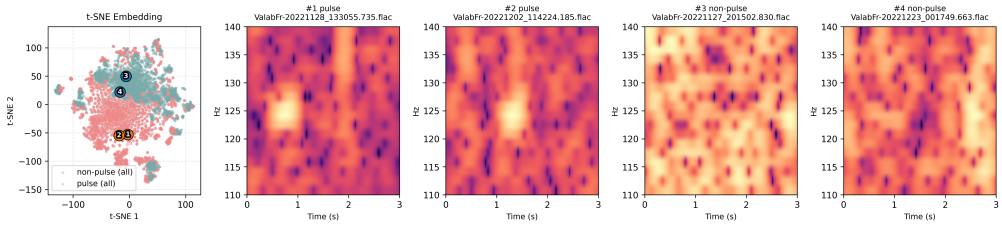

Figure 6: Left: t-SNE projection of frozen SSL embeddings. Right: linear spectrograms of examples

F Ablation Study

F.1 Normalization Method

To verify the effect of normalization on amplitude preservation, we conducted an ablation study by replacing our proposed BRN with alternative normalization schemes, including Group Normalization (with group sizes 1 and 128), Layer Normalization, Batch-Instance Normalization (BIN), RMS Normalization, and Batch Normalization.

Each variant was trained using the same CPC framework and evaluated on the validation set. We report CPC loss and CPC accuracy as indicators of pretraining quality, and further assess downstream utility by extracting embeddings and applying linear probe classifiers (logistic regression with PCA) and k-NN. Each experiment is done with 20 epochs on 4 $\times$ 3090 RTX GPUs for about 2 hours. All ablation results are reported as mean $\pm$ SEM over 3 independent runs due to the high computational cost of training.

Normalization	F1	AUROC	Precision	Recall
Group Norm (gs=1)	0.638 ± 0.037	0.742 ± 0.011	0.664 ± 0.024	0.622 ± 0.039
Group Norm (gs=128)	0.660 ± 0.023	0.725 ± 0.037	0.674 ± 0.032	0.655 ± 0.020
Layer Norm	0.791 ± 0.016	0.861 ± 0.012	0.865 ± 0.033	0.737 ± 0.020
BIN (Batch+Instance)	0.647 ± 0.006	0.736 ± 0.008	0.676 ± 0.004	0.631 ± 0.009
RMS Norm	0.847 ± 0.014	0.908 ± 0.009	0.899 ± 0.006	0.807 ± 0.017
Batch Norm	0.856 ± 0.003	0.913 ± 0.003	0.898 ± 0.012	$\mathbf{0.822 \pm 0.002}$
BRN (ours)	$\mathbf{0.856 \pm 0.002}$	$\mathbf{0.914 \pm 0.002}$	$\mathbf{0.900 \pm 0.002}$	0.813 ± 0.008

Table 5: Ablation study of different normalization schemes on CPC encoder. Results are reported as mean $\pm$ SEM over three runs. Metrics are computed via linear probe evaluation on the validation set.

Normalization	KNN	CPC Acc.	CPC Loss
Group Norm (gs=1)	0.642 ± 0.026	0.58 ± 0.21	1.83 ± 0.81
Group Norm (gs=128)	0.646 ± 0.028	$\mathbf{0.87 \pm 0.05}$	$\mathbf{0.62 \pm 0.09}$
Layer Norm	0.714 ± 0.007	0.40 ± 0.01	4.02 ± 0.03
BIN (Batch+Instance)	0.646 ± 0.003	0.86 ± 0.06	0.68 ± 0.19
RMS Norm	0.728 ± 0.005	0.41 ± 0.01	4.02 ± 0.06
Batch Norm	0.732 ± 0.003	0.41 ± 0.00	3.87 ± 0.20
BRN (ours)	$\mathbf{0.737 \pm 0.008}$	0.40 ± 0.01	3.81 ± 0.25

Table 6: Additional ablation results on KNN evaluation and CPC training metrics. Results are reported as mean $\pm$ SEM over three runs.

The ablation results highlight the importance of preserving amplitude cues in fin-whale pulse detection. Commonly used schemes such as LayerNorm, GroupNorm, and BIN tend to normalize away absolute energy, resulting in embeddings that are less informative for downstream classifiers, as reflected by relatively lower F1 scores and AUROC values. By contrast, RMSNorm, which does not subtract the mean, retains more amplitude information and achieves stronger separation metrics. BatchNorm and our proposed BRN yield comparable downstream results, but BatchNorm is well known to

suffer from instability with small batch sizes and poor generalization under distribution shift. BRN alleviates these issues by interpolating between BN and RMSNorm, allowing it to capture amplitude-sensitive features without over-reliance on batch statistics. This design leads to more robust amplitude preservation and consistent performance across training conditions.

F.2 SincNet vs. Standard Convolution Front End

To assess the contribution of the Sinc-based convolutional layer, we replace it with a standard learnable convolutional front-end while keeping BRN as the normalization method. The experiment settings are same with settings in Appendix F.1. As shown in Table 7, the SincNet front-end substantially improves downstream linear probe performance across all metrics. In particular, F1 score increases from 0.784 to 0.856, and AUROC improves from 0.868 to 0.914, while precision and recall also show significant gains. By constraining filters to band-pass responses, SincNet effectively emphasizes frequency bands that carry biologically relevant pulse energy while suppressing irrelevant noise. These results confirm that SincNet not only enhances the informativeness of the learned representations but also reduces the impact of spurious noise, yielding embeddings better suited for downstream pulse detection tasks than traditional convolutional layers.

Metric	Conv (plain)	SincNet (ours)
F1	0.784 ± 0.004	$\mathbf{0.856 \pm 0.001}$
AUROC	0.868 ± 0.003	$\mathbf{0.914 \pm 0.001}$
Precision	0.849 ± 0.007	$\mathbf{0.900 \pm 0.001}$
Recall	0.742 ± 0.003	$\mathbf{0.813 \pm 0.006}$
KNN	0.693 ± 0.003	$\mathbf{0.737 \pm 0.005}$
CPC Acc.	0.387 ± 0.009	$\mathbf{0.402 \pm 0.008}$
CPC Loss	$\mathbf{3.679 \pm 0.103}$	3.807 ± 0.143

Table 7: Replacing the Sinc-based front-end with a standard convolutional front-end. Results are mean $\pm$ SEM over three runs on the validation set, following the same evaluation protocol as Section F.1.

G Model Architecture

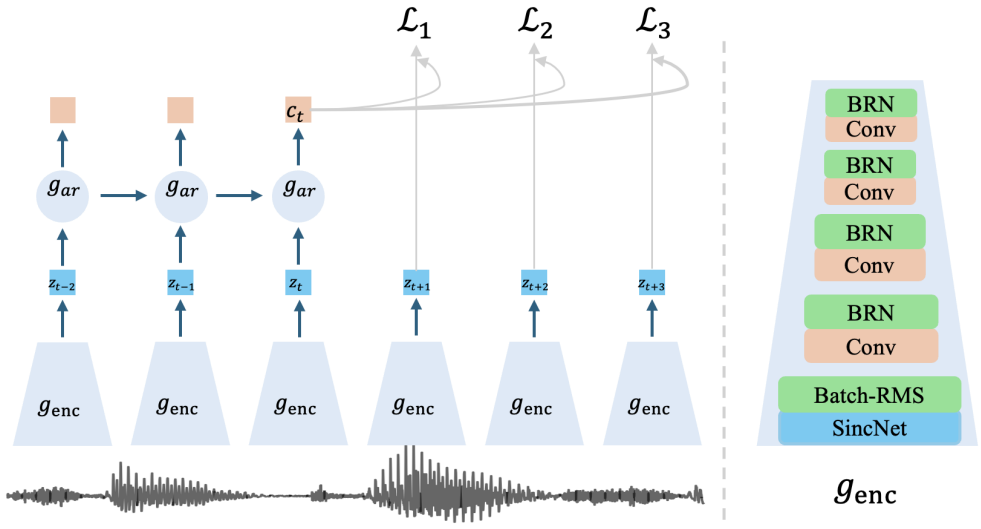

Figure 7: The architecture of the Self Supervised model. Left: CPC model with encoder g_{enc} and autoregressive model g_{ar}. Right: The structure of g_{enc}.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the contributions—first applying SSL to fin-whale detection and showing gains in limited labeled data, low SNR and cross-site settings—and these claims are directly supported by the experiments in Section 4.2 and Appendix D.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: In Section 4.2, we found that in high-label regimes SSL underperforms supervised models, likely because a frozen encoder cannot benefit from more labels. This limitation and possible improvements are discussed in Section 5.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: This paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The Mediterranean dataset split follows Best et al. (2022), while the Seglvik dataset uses YOLOv5 detections for training and manually validated annotations for validation and test. Model architectures are detailed in Section 3.1–3.2, with hyperparameters in Appendix B. Together with planned dataset and code release, these details ensure our experiments are reproducible.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Anonymized links to datasets and code are provided in Appendix A.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.

- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All the training and test details can be found in Section 4.1 and Appendix B.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report mean $\pm$ SEM performance across 10 random seeds for main experiments and mean $\pm$ SEM across 3 runs for ablations which can be found in Section 4.2 and Appendix F.1 & F.2. Fig.1 & 2 also visualize uncertainty with shaded areas representing the SEM. Ablations were run three times due to computational cost.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: The computer resources for each experiment, including Ablation, has been reported in Appendix B & F.1.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The study uses publicly available or ethically collected bioacoustic whale datasets and fully complies with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: The work focuses on whale bioacoustics and does not have direct societal impact on humans. Its scope is limited to ecological monitoring research.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.

- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The work does not involve high risk models and data. The datasets consist of whale bioacoustic recordings with no sensitive or personal information.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The FinWhaleSong dataset Best et al. [2022] we used is released under a CC-BY 4.0 license, and YOLOv5 is AGPL-3.0; We cite and use all assets strictly within their license terms for academic research.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We introduce a new annotated dataset from Seglvik Fjords recordings and code for the SSL model. Documentation is provided in Section 3 and Appendix A, and anonymized links to the assets are included in the Appendix A.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.