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

Adam Chareyre[1,4]*, **Haodong Zhang**[2]*, **Shuwen Ge**[3], **Randall Balestriero**[2,4], **Hervé Glotin**[1,4]
[1]Université de Toulon, Aix Marseille Univ, CNRS, LIS DYNI, Toulon, France
[2]Dpt of Computer Science, Brown University, USA [3]Xi'an University of Technology, China
[4]CIAN, Int. Center of AI for Natural Acoustics, https://cian.univ-tln.fr

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 we collected in an arctic fjord in Norway, and in the Mediterranean Sea, SSL models outperform supervised Transformers Encoder in low-label (respectively 88.5% and 68.6% f1-score for 0.1% of the training set size) and low SNR regimes (respectively 87.4% and 81.3% f1-score for SNR ≤ -5 with the training set) 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], Girardet et al. [2025]. 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 areas: in arctic fjord Glotin et al. [2025], and in Mediterranean Sea Glotin et al. [2023], Laran et al. [2009].

*Equal contribution. Emails: adam.chareyre.1999@gmail.com; haodong_zhang@brown.edu; 3230412084@stu.xaut.edu.cn; randall_balestriero@brown.edu; glotin@univ-tln.fr.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI for Non-Human Animal Communication.

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 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 on our recording device QHB Barchasz et al. [2020]. 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 encoder 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. Detailed pipeline of the Supervised model is shown in Appendix G.1.

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.

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 $\{z_j\}_{j=1}^{N}$.

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

with $\mathrm{sim}(u, v) = u^\top v$ 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.2.

3.3 Datasets

We evaluate on two fin-whale datasets. The first is from our **Seglvik arctic Fjord (Norway)** shored recording device Glotin et al. [2025], Girardet et al. [2025]. It comprises over 1500 hours of recordings collected in a noisy Arctic environment with seasonal whale aggregations. Training labels were obtained through YOLOv12 Tian et al. [2025], totaling 107,741 annotations, all confidence levels considered, while validation and test sets were fully manually curated, totaling 433 and 399 annotations. Each sample is represented as a 3-s segment centered on the 125 Hz pulse band. We apply data augmentation by performing time-shift operations around each pulse, generating two additional samples per pulse. This augmentation is applied to both the validation and test sets to reduce model variance and improve the consistency and stability of the results. The second, the **Mediterranean Sea** data, is from our sonobuoy BOMBYX1 Glotin et al. [2023], Poupard et al. [2022], plus the Boussole sonobuoy Laran et al. [2009], yielding to our FinWhaleSong dataset Best et al. [2022], of nearly 3,700 pulses at 20 Hz, annotated by an active learning process with a CNN.

We then have created 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. They are all available at `https://cian.lis-lab.fr/cianscape`.

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 Fjord** 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 ($\tau \in -4, 0, 1, 2, 4, 8$ dB) while evaluation remains on the full testset. Additional implementation details, including optimizer settings and model sizes, are provided in Appendix B.

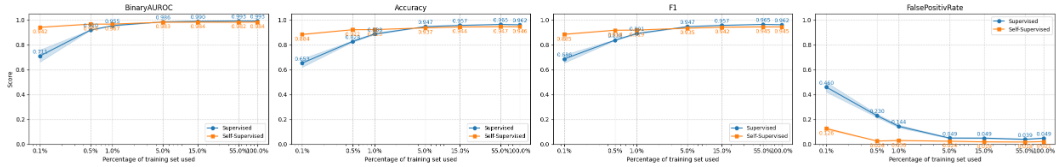

Figure 1: Performance of supervised method and SSL method under varying training set sizes on Seglvik Fjord dataset, demonstrating the robustness of our SSL compared to Supervised.

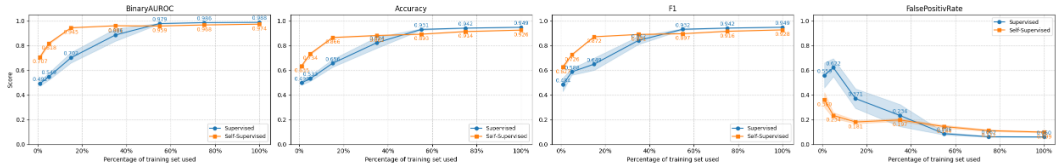

Figure 2: Performance of supervised method and SSL method under varying training set sizes on Mediterranean Sea.

4.2 Result & Analysis

On the **Seglvik Fjord** dataset, the supervised Transformer Encoder and SSL-pretrained encoder perform comparably with full supervision, as shown in Table 1 and 2. The **Mediterranean Sea** dataset's results are shown in Table 3 and 4 in Appendix C.

Model	Acc	AUROC	FPR	F1
Supervised	0.962 ± 0.001	0.993 ± 0.000	0.049 ± 0.004	0.962 ± 0.001
SSL	0.949 ± 0.001	0.986 ± 0.000	0.029 ± 0.002	0.948 ± 0.001

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

Model	Recall	Precision	Loss
Supervised	0.972 ± 0.003	0.953 ± 0.004	0.350 ± 0.001
SSL	0.928 ± 0.002	0.969 ± 0.002	0.365 ± 0.000

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

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 Fjord testset into the frozen pretrained model, extract embeddings with RMS-weighted pooling, and

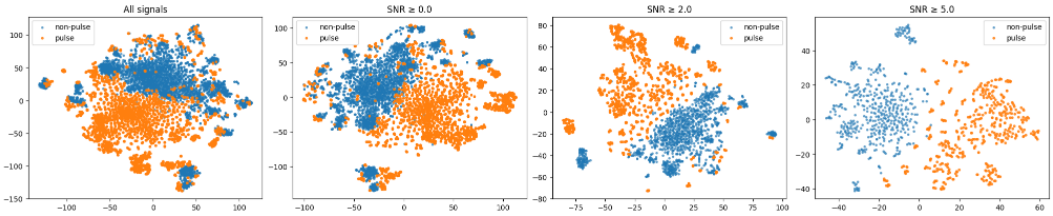

Figure 3: **t-SNE embeddings of the SSL model at different SNR thresholds with the Seglvik dataset.** Left to right: all signals, SNR ≥ 0, ≥ 2, and ≥ 5 dB.

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.

Acknowledgments

We first thank Gianni Pavan for providing the Magnaghi data, which gave us the first reference of Mediterranean fin whale pulses, necessary to start to learn the analysis of the rest of the data (recordings made by CIBRA, Univ. of Pavia, with a sonobuoy from the IT Navy Magnaghi in the Saclantcent SIRENA99 sea trial).

This research is partly granted by ANR-20-CHIA-0014-01 national Chair in Artificial Intelligence for Bioacoustics (ADSIL) (H.G.) AID DGA ANR, and by Biodiversa+, the European Biodiversity Partnership under the 2021-2022 BiodivProtect co-funded by the European Commission (GA N°101052342): EUROPAM 2023-2026 Biodiversa2021-488 (H.G.). We thank CIAN for the internship and ing. positions of Adam Chareyre who conducted this paper until sept. 2025.

For BOMBYX1, we thank Osean SAS for recorders, Park national of Port-Cros, PMS SAS for logistics, Prefecture Maritime de la Méditerranée for scientific permit. We thank G. Rougier for his help in the mooring. BOMBYX1 was designed by H.G. and co-funded by its chair at Institut Universitaire de France, TPM, CG83, Univ. of Toulon, Pole INPS, LIS CNRS, and Engie Fondation. We thank MI CNRS MASTODONS SABIOD and Region PACA and GIAS Marittimo for bigdata storage. We thank The Pelagos Sanctuary which granted the studies on BOUSSOLE and BOMBYX.

For the arctic recordings, we thank MITI CNRS ADAPREDAT FJORD3D missions, and J. Guiderdoni at ValhallaB for great help in the installation of our recording station in Arctic, Seglvik. We thank Marion Poupard and Jean-Marc Prevot for the installation of this acoustic array with H.G. Thank to the Toulon Provence Méditerranée, APRI UTLN, ANR ULPCOCHLEA (ANR-21-CE04-0020) (H.G.) and EUROPAM Biodiversa (H.G.) for support to our Arctic expeditions 2022-23.

We also thank Professor Sébastien Paris for his supervision and guidance throughout this work.

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.

Valentin Barchasz, Valentin Gies, Sebastian Marzetti, and Hervé Glotin. A novel low-power high speed accurate and precise DAQ with embedded artificial intelligence for long term biodiversity survey. In *e-Forum Acusticum*, pages 3217–3224, Lyon, France, December 2020. doi: 10.48465/fa.2020.0875. URL `https://hal.science/hal-03230835`.

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.

Justine Girardet, Hervé Glotin, Stéphane Chavin, Marion Poupard, Julie Guiderdoni, and Véronique Sarano. Arctic diel and circadian acoustic pattern of orcas, fin, and humpback whales revealed by deep learning from two months of continuous recordings. *Ecological Informatics*, (to appear), 2025.

Hervé Glotin, Stéphane Chavin, Justine Girardet, Pascale Giraudet, Paul Best, Maxence Ferrari, Véronique Sarano, François Sarano, Pierre Mahé, Valentin Barchasz, Valentin Gies, Nicolas Deloustal, Fabien de Varenne, Julie Patris, Jean-Marc Prévot, Olivier Philippe, Franck Hieramente, Vincent Bertin, Paschal Coyle, and Sébastien Paris. Bilan d'une décennie d'observations des grands cétacés en milieu anthropisé Nord Pelagos: BOMBYX, KM3NeT, et antennes mobiles Sphyrna et WhaleWay. Technical report, CIAN, October 2023. URL `https://univ-tln.hal.science/hal-04939839`.

Hervé Glotin, Jean-Louis Etienne, Stéphane Jespers, Justine Girardet, Pascale Giraudet, Valentin Gies, Véronique Sarano, François Sarano, Jean-Marc Prévot, Nathalie D'alvise, Sébastian Marzetti, Valentin Barchasz, Marion Poupard, Maxence Ferrari, Lionel Camus, Pierre Priou, Malik Chami, Julie Guiderdoni, and Sébastien Paris. Orca, Humpback, Fin and Sperm whales all together into an anthropized Arctic Fjord: the 2021 to 2024 CIAN Expeditions. Technical report, CIAN, July 2025. URL `https://univ-tln.hal.science/hal-05142566`.

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.

Sophie Laran, Manuel Castellote, Frederic Caudal, and Hervé Glotin. Suivi acoustique des cétacés au nord du sanctuaire PELAGOS. Technical report, Pelagos France, CNRS, Univ. Toulon, 2009. URL `https://pelagos-sanctuary.org/wp-content/uploads/2024/11/42.-Laran-et-al-2009_Suivi-acoustique-des-cetaces-au-nord-du-Sanctuaire-Pelagos.-2007-2009.pdf`.

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.

Marion Poupard, Maxence Ferrari, Paul Best, and Hervé Glotin. Passive acoustic monitoring of sperm whales and anthropogenic noise using stereophonic recordings in the mediterranean sea, north west pelagos sanctuary. *Scientific reports*, 12(1):2007, 2022.

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.

Yunjie Tian, Qixiang Ye, and David Doermann. Yolov12: Attention-centric real-time object detectors. *arXiv preprint arXiv:2502.12524*, 2025.

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 Fjord (Norway). Our Seglvik Fjord dataset was collected in northern Norway, designed and placed during our Arctic expedition in **Seglvik arctic Fjord (Norway)** Glotin et al. [2025], Girardet et al. [2025]. It 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 831 fin-whale pulses and automatically generated 831 negative annotations by selecting random portions of the signal outside the positive annotations. A YOLOv12 detector was then trained and applied to the entire 1555 hours of recordings, producing 107,741 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 107,741 to 46,377 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 831 manually validated 125 Hz fin whale pulses, and the validation set contains 831 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.

To convert the raw signal into a time–frequency representation, we computed a short-time Fourier transform (STFT) using a 2048-sample FFT and a 2048-sample Hann window, with a hop size of 120 samples. A band-pass filter was then applied to retain frequencies within the [100 Hz, 150 Hz] range.

Mediterranean Sea (Bombyx and Boussole). We also evaluate on the Fin Whale Songs that we collected in the Northwestern Mediterranean from in 2008 by our Boussole sonobuoy device Laran et al. [2009], and our BOMBYX sonobuoy in 2015-2018 Glotin et al. [2023], Poupard et al. [2022]. We pooled in the FinWhaleSong dataset Best et al. [2022]. 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. We only kept annotations with a confidence level above 0.9, leaving us with 1,231 pulse annotations. Since the public version is not pre-partitioned, we 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 pulse once along the time axis. We also filtered the signal with the 15-35 Hz frequency band.

To convert the raw signal into a time–frequency representation, we computed a short-time Fourier transform (STFT) using a 256-sample FFT and a 256-sample Hann window, with a hop size of 6 samples. A band-pass filter was then applied to retain frequencies within the [15 Hz, 35 Hz] range.

B Experimental Setup

Supervised training. For supervised models, we use the AdamW optimizer with a weight decay of 1×10^{-5}, a OneCycleLR scheduler, and a binary cross-entropy loss.

On the *Mediterranean Sea* dataset, the model (60k parameters) is trained for 20 epochs with a batch size of 32, including data augmentation through time-shifting. Training completes in 1 minute on a single A40 GPU when considering all SNR conditions and applying a confidence threshold of ≥ 0.9. As in the original setup, an additional computation of the optimal threshold is performed on

the validation set. This low number of parameters ensures a fair comparison with prior CNN-based baselines Best et al. [2022].

On the *Seglvik* dataset, the model (84k parameters) is trained for 10 epochs with a batch size of 32, combining all SNR levels and using a confidence threshold of ≥ 0.9. The full training process requires 4 minutes on an A40.

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 3 & Table 4 present results on the Mediterranean dataset. Models were trained with the complete training dataset whose annotations have a confidence level ≥ 0.9, and evaluation was performed using an automatically selected optimal threshold strategy based on the results of the validation set.

Model	Acc	AUROC	FPR	F1
Supervised	0.951 ± 0.002	0.988 ± 0.001	0.059 ± 0.003	0.951 ± 0.003
SSL	0.927 ± 0.002	0.974 ± 0.001	0.098 ± 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

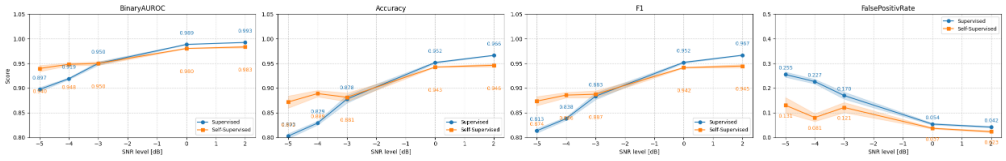

Figure 4: Performance of supervised and SSL-based models on Seglvik when training is restricted to subsets with SNR $\leq \tau$, while testing is performed on the full test set. Both methods benefit from increasing SNR thresholds, but at the lowest SNR levels (e.g., $\tau = -5$ dB) the SSL model achieves better performance.

To directly assess robustness to noisy training data, we construct training sets on the Seglvik dataset by applying a maximum SNR threshold and using only samples with SNR $\leq \tau$, where $\tau \in \{-5, -4, -3, 0, 2\}$ dB. The test set remains unchanged and covers the full distribution of SNR levels. Each configuration is repeated with ten random seeds for statistical reliability.

As shown in Fig. 4, performance decreases for both approaches when the training SNR threshold is reduced, but the degradation is consistently milder for the SSL model. In the most adverse setting ($\tau = -5$ dB), SSL attains about 7 percentage points higher accuracy and F1 than the supervised model, and reduces the false positive rate by roughly a factor of two. As the SNR threshold increases,

the gap between SSL and supervised models narrows and almost vanishes for $\tau \geq -3$ dB. These results indicate that SSL can learn useful and more robust representations even when supervision is provided only on extremely noisy signals, while supervised training is more sensitive to such adverse conditions.

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.

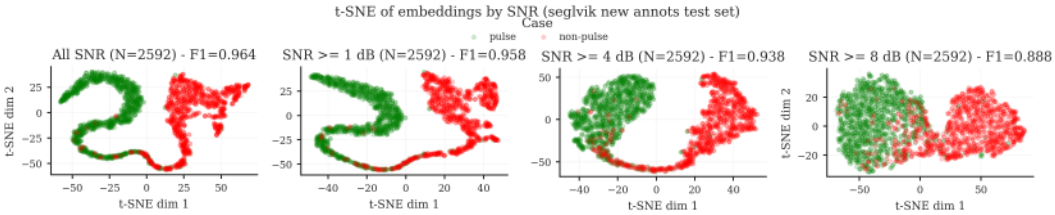

Figure 5: **t-SNE embeddings of the supervised model on the test set at different SNR thresholds with the Seglvik training set.** Left to right: all signal SNRs, SNR ≥ 1, ≥ 4, and ≥ 8 dB.

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.

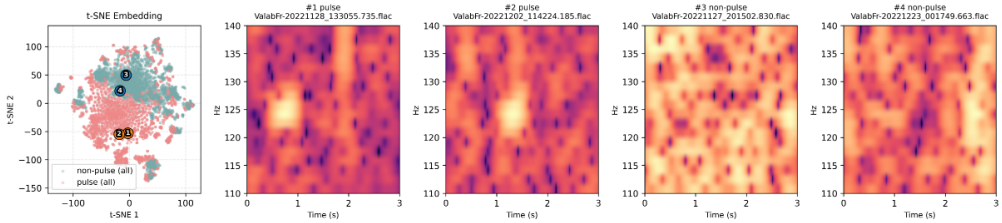

Figure 6: Left: t-SNE projection of frozen SSL embeddings. Right: linear spectrograms of examples

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

G.1 Supervised Architecture

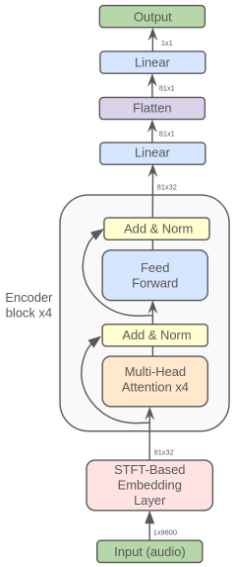

Figure 7: The supervised model pipeline: containing the STFT transformation followed by the supervised model architecture, a refined transformer encoder model.

G.2 SSL Architecture

H Fairness Analysis: Fixed Spectrogram Front-End vs. Learnable Sinc-Based Front-End

In the supervised baseline, the front-end is a fixed STFT-based spectrogram, whereas the SSL model employs a learnable SincNet filterbank. Since the SSL front-end is trainable, it is natural to ask whether this could lead to an unfair advantage. We analyze this point in detail.

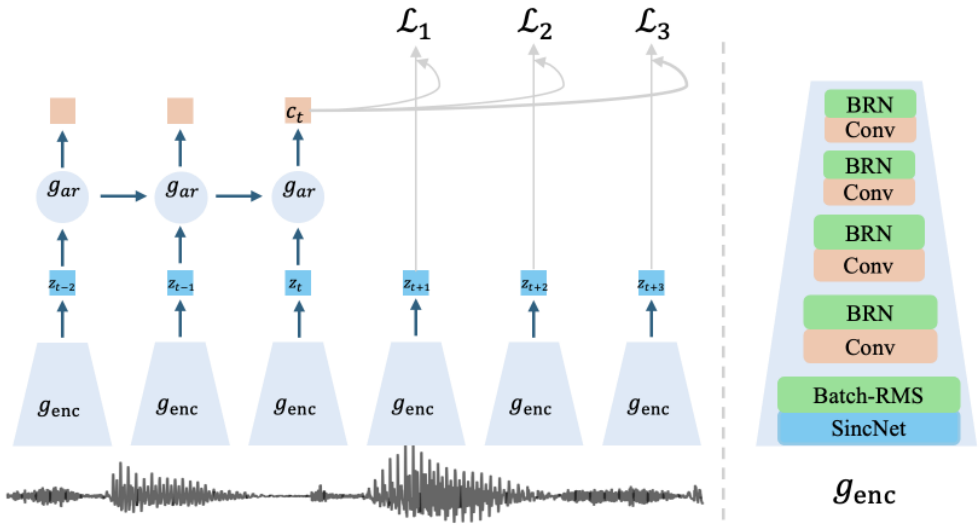

Figure 8: The architecture of the Self Supervised model. Left: CPC model with encoder g_{enc} and autoregressive model g_{ar}. Right: The structure of g_{enc}.

Fixed STFT does not disadvantage the supervised baseline. A fixed STFT provides a deterministic and physically interpretable time–frequency representation. For low-frequency whale vocalizations, such representations form a well-established and acoustically meaningful front-end. The predefined time–frequency resolution acts as a strong inductive bias, stabilizing optimization and enabling supervised models to converge rapidly, particularly when labeled data are scarce or noisy. In this sense, STFT does not weaken the supervised baseline; instead, it provides a reliable signal decomposition that the model does not need to learn.

Learnable SincNet introduces a more difficult optimization problem. In contrast, the SSL model must jointly learn the low and high cutoff frequencies of each filter, together with the latent representation required for contrastive prediction. This substantially increases the complexity of the optimization landscape. Unlike a pre-computed spectrogram, the Sinc front-end receives no explicit time–frequency decomposition and must discover appropriate band-pass filters purely through self-supervised objectives. This typically requires more computing resources, longer training periods, and larger datasets before converging to meaningful filters.

The difference in front-end design does not inherently favor SSL. The supervised baseline benefits from a stable and well-understood representation with strong acoustic priors, whereas the SSL model bears the additional burden of learning the filterbank itself. Therefore, the performance gains observed for SSL cannot be attributed to having a "stronger" or more flexible front-end; they instead reflect the contribution of representation learning on large amounts of unlabeled data.

I Resources

All datasets and code used in this study are publicly available. The repository provides the Seglvik dataset, the Mediterranean dataset, and the full codebase required to reproduce all experiments presented in this paper. The resources can be accessed at: `https://huggingface.co/datasets/CIANLabxBROWNUniv/fin_whale_1`. The commit number corresponding to the version of this paper is: 9734044922d5590cda9b3dc26b734b8206788c2e. We also make available the raw, unsplit data and the corresponding annotations at the following URL: `https://cian.lis-lab.fr/cianscape`.