doi:10.26443/seismica.v5i1.1821

Semantic segmentation for feature detection in ocean bottom seismometer data

Alex A. Saoulis * [1,2], **Afonso Loureiro** [3,4], **Maria Tsekhmistrenko** [1,5], **Ana M.G. Ferreira** [1]

[1]Department of Earth Sciences, University College London, London, United Kingdom, [2]Department of Physics & Astronomy, University College London, London, United Kingdom, [3]ARDITI - Regional Agency for the Development of Research, Technology and Innovation, Funchal, Portugal, [4]University of Lisbon, Institute D. Luiz, Lisbon, Portugal , [5]ERP, Earth Rover Program, London, United Kingdom

Author contributions: *Conceptualization*: A.A.S., A.L., M.T., A.M.G.F.. *Methodology*: A.A.S., A.L., M.T., A.M.G.F.. *Software*: A.A.S.. *Writing - Original draft*: A.A.S., A.L.. *Writing - Review & Editing*: A.A.S., A.L., M.T., A.M.G.F.. *Supervision*: A.L., M.T., A.M.G.F..

Abstract We introduce semantic segmentation for time–frequency representations of seismic data, enabling pixel-level detection and characterisation of signals. We curate a small, manually annotated dataset of 500 spectrograms containing a range of feature classes in the 1 to 50 Hz frequency band from a single ocean bottom seismometer (OBS) from the UPFLOW array in the mid-Atlantic region. We explore several machine learning (ML) training techniques that are specialised for low-data training regimes, and compare their performances for two feature classes (instrument resonances and blue whale calls). We find that a synthetic pretraining step significantly improves performance relative to semi-supervised approaches and finetuning an off-the-shelf model, with a $\sim$5% improvement in performance for well-represented features, and an improvement of over 100% for rare features. Despite the small dataset, our method can be utilised to accurately and efficiently segment spectrogram data across 43 OBSs from the large-scale UPFLOW array, as well as data from previous OBS deployments. We next investigate a range of applications for the trained segmentation models. We demonstrate that our ML algorithm identifies current-induced instrument resonances accurately enough to extract a tidal signal. In addition, it reliably detects blue whale calls across the entire UPFLOW array, and it even enables automated tracking of individual whales detected simultaneously at multiple OBSs.

Production Editor:
Yen Joe Tan
Handling Editor:
Lise Retailleau
Copy & Layout Editor:
Anant Hariharan

Signed reviewer(s):
Vaibhav Vijay Ingale

Received:
July 8, 2025
Accepted:
December 16, 2025
Published:
February 25, 2026

1 Introduction

Ocean-bottom seismology offers a unique glimpse into a diverse set of processes at the seafloor. Typically, the primary motivation of ocean-bottom seismometer (OBS) experiments is to improve seismic data coverage and imaging quality in the oceans, which are some of the most under-instrumented regions on Earth. However, the exposure and proximity of OBSs to a wide variety of noise sources often represent a problem for typical seismological studies.

Noise for one scientific objective, however, may represent signal for another. Natural and anthropogenic noise sources on land stations and OBSs have been used to glean insights into a wide range of processes. Cryosphere and weather-related phenomena such as ice calving, sea-ice cover, and ocean storms have been investigated (O'Neel et al., 2007; Walter et al., 2010; Nettles and Ekström, 2010; Aster et al., 2008; Anthony et al., 2014; Bromirski et al., 2005; Kedar et al., 2008; Koper and Burlacu, 2015; Davy et al., 2014; Barruol et al., 2015; Gualtieri et al., 2018). Anthropogenic signals include ship tracks (Pakhomov and Goldburt, 2006; Trabattoni et al., 2023). Additional studies have explored biological activity (Gaspà Rebull et al., 2006; Dréo et al., 2019; Pereira et al., 2020) and even tracked ocean currents using instrument noise (Stähler et al., 2018; Essing et al., 2021; Corela et al., 2023; Godin et al., 2024; Tan et al., 2025). Identifying such signals heavily relies on manual inspection and expert-informed heuristic methods, which are time-consuming and difficult to scale. In this study, we explore how modern machine learning (ML) methods can be used to detect and track a wide variety of non-seismic signals efficiently, with relatively limited manually annotated datasets.

Most of the previous work using OBS data has focused on noise estimation and removal. Various algorithms have been developed to remove and decrease OBS noise, such as compliance noise (Crawford and Webb, 2000; Bell et al., 2015). A range of approaches utilising algorithmic noise estimation and filtering techniques have also been explored (e.g., Mousavi and Langston, 2017; Negi et al., 2021; An et al., 2022; Zali et al., 2023). ML has also found significant success in improving the signal-to-noise ratio for seismological data analyses through denoising (e.g., Zhu et al., 2019; Yu et al., 2019; Dahmen et al., 2022; Chen et al., 2024). Generally, such approaches suppress noise without distinguishing between its diverse physical sources. This study instead utilises ML to automatically detect and characterise these signals in seismic data.

One signal class of interest here are instrument vibrations detected by the seismometer. On the seafloor, OBSs are exposed to oceanic currents. They are obstacles around which water must flow, creating vortices that interact with components of the instrument.

*Corresponding author: a.saoulis@ucl.ac.uk

The periodic lift and drag effects, coupled with the mechanical strumming of components, lead to the production of vibrations that propagate through the frame of the instrument and are recorded by the seismometer (e.g., Corela et al., 2023). In certain conditions, the frequency at which vortices are shed from a specific component matches its eigenfrequency, and constructive interference leads to the generation of overtones, or resonances, which are recorded by the seismometer (Griffin, 1985; Trehu, 1985b; Stähler et al., 2018).

Another class of signals often recorded by OBSs are baleen whale vocalisations, particularly those of blue whales (Dunn and Hernandez, 2009; Brodie and Dunn, 2015; Dréo et al., 2019; Wilcock and Hilmo, 2021) and fin whales (McDonald et al., 1995; Gaspà Rebull et al., 2006; Wilcock, 2012; Matias and Harris, 2015; Pereira et al., 2020). These are well-suited for detection by seafloor seismic arrays due to the low frequency (< 50 Hz) and relatively high amplitude of the vocalisations (> 180 dB relative to 1 µPa at 1 m; Širović et al., 2004; Miller et al., 2021b). Automated detection and characterisation of whale vocalisations is an active area of research, using both seismic instruments and specialised passive acoustic monitoring instruments (Baumgartner and Mussoline, 2011; Miller et al., 2021a; Rasmussen and Širović, 2021; Allen et al., 2021; Plourde and Nedimović, 2022; Goodwin et al., 2022; Stowell, 2022; Miller et al., 2023; Napoli and White, 2023; Cotillard et al., 2024). This study focuses on the detection of blue whale vocalisation.

ML has seen wide success in seismology for signal detection and arrival picking (Mousavi and Beroza, 2022, 2023). These methods most often operate directly in the time domain, performing classification of fixed-length windows or 1-D *segmentation*, where the probability of an arrival is predicted for each time step (e.g., Zhu and Beroza, 2018; Mousavi et al., 2020; Münchmeyer et al., 2022; Woollam et al., 2022; Bornstein et al., 2024). A smaller number of studies instead use time–frequency representations, but these approaches either classify entire spectrogram segments (Nakano et al., 2019; Stepnov et al., 2021; Shakeel et al., 2022; Xi et al., 2024; Tan et al., 2024; Si et al., 2024) or perform 1-D first-arrival segmentation using time-frequency images (Mousavi et al., 2019b; Saad et al., 2021; Choi et al., 2024; Peng et al., 2025).

Here, we introduce 2-D semantic segmentation for time–frequency representations of seismic data. This approach assigns a class label to every pixel of the spectrogram (for a review, see Minaee et al., 2021), which has the unique advantage of enabling detailed characterisation of detected signals, including their temporal extent, spectral bandwidth, and even energy content. Our formulation parallels semantic segmentation in natural images (Minaee et al., 2021; Peláez-Vegas et al., 2023; Zhou et al., 2024), enabling us to draw on mature computer vision methods. These techniques have now proven effective in audio (Jansson et al., 2017; Venkatesh et al., 2022), and have even seen (very limited) adoption in bioacoustics (Jin et al., 2022).

In addition, the significant successes of ML in seismology can be attributed in part to the very large, openly available seismic arrival datasets (e.g., ISC-GEM; Storchak et al., 2013 and STEAD; Mousavi et al., 2019a), which allow for the supervised training of deep learning ML models. However, in settings where no large datasets exist, adopting these ML techniques remains challenging (e.g., in understudied signal classes and new regions; Lapins et al., 2021; Jiang et al., 2021; Wang et al., 2023; Koper et al., 2024; Zhong and Tan, 2024; Zhu et al., 2023, or different instrument classes such as OBSs; Bornstein et al., 2024; Niksejel and Zhang, 2024). Therefore, this study investigates strategies to improve model performance under limited annotated training data.

In this study, we use data from the recent *Upward mantle flow from novel seismic observations* (UPFLOW) (Tsekhmistrenko et al., 2025) OBS array in the Azores-Madeira-Canary Islands region (Atlantic Ocean). We develop and evaluate several approaches for training semantic segmentation models to detect and classify different signals recorded on OBS data. We focus on identifying two signals: instrument resonances in the 4 Hz to 30 Hz range and ∼17 Hz blue whale vocalisations.

This paper is structured as follows. Section 2 presents the data used in this study and introduces key background information on OBS noise relevant to this study. Section 3 then presents the methodology underpinning this work, covering the ML-based techniques, as well as the modified training strategies utilised to deal with small annotated datasets. The results are given in Section 4, exploring model performance on instrument reverberation and blue whale vocalisation detections. Section 4 also presents a range of downstream proof-of-concept applications to validate the methodology. Finally, Section 5 presents the discussion and conclusions.

2 Data

2.1 The UPFLOW array

Fifty free-fall broadband OBSs were deployed and 49 were recovered as part of the largest passive seafloor experiment in the Atlantic Ocean so far – the UPFLOW experiment in the Azores-Madeira-Canary Islands region between June 2021 and August 2022 (Ferreira, 2024; Tsekhmistrenko et al., 2025). The UPFLOW array, for which typical inter-station distances were 110-160 km, is shown in Fig. 1. All OBSs were equipped with hydrophones; 49 had three-component broadband seismometers, which were mostly Trillium Compact seismometers (with a corner period of T=120 s), and one OBS (UP17) included a Silicon Audio accelerometer. The sampling rate of the OBSs used in this study was either 100 Hz or 250 Hz. Following comprehensive data quality analyses, Tsekhmistrenko et al. (2025) identified three malfunctioning OBSs - UP10, UP19 and UP23 - likely due to broken sensors and/or damaged components (e.g., damaged pins on connectors), which are not used in this study. Furthermore, three OBSs were prototypes that either had data issues due to firmware failures (UP07, UP24) or had an accelerometer, which was not useful for this study (UP17)(Tsekhmistrenko et al., 2025) and

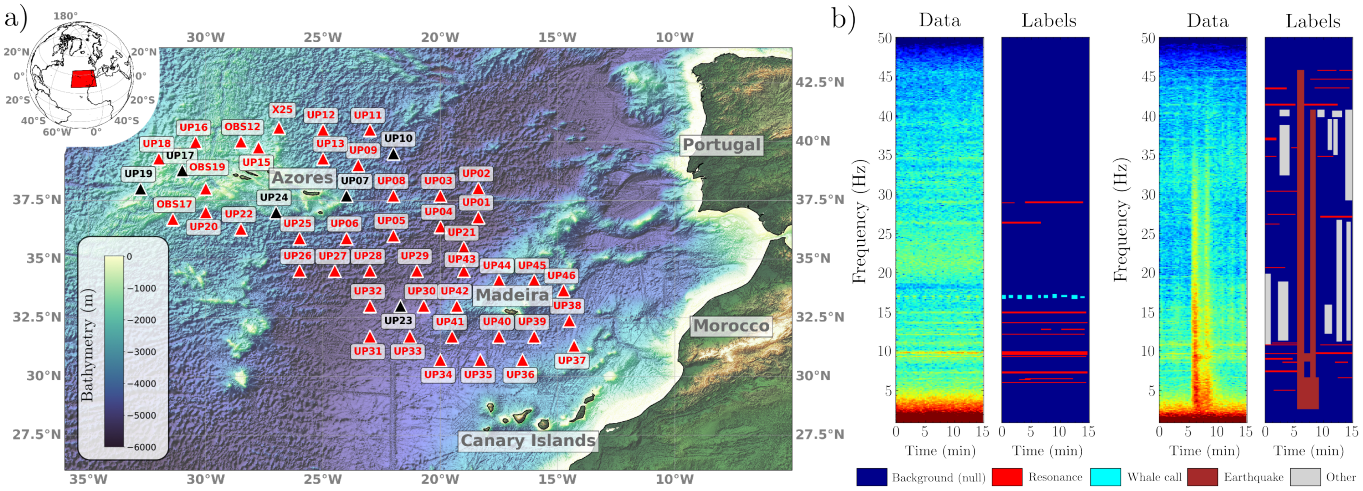

Figure 1 a) The UPFLOW OBS array located in the Azores-Madeira-Canaries region. Each triangular marker denotes the location of the 49 OBSs that were recovered: red markers denote stations whose data were used in this study, while black markers denote stations with data issues. b) Two random examples of spectrograms from the UPFLOW UP05 station, alongside the manually annotated regions that corresponds to data features of interest. This work utilises ML techniques to automatically detect feature regions corresponding to instrument resonances (red) and blue whale calls (light blue).

hence are not used either. For all other stations, it was found that overall, the data quality was mostly high. For example, vertical component data showed lower long-period noise levels than in previous experiments. In addition, the quality of the horizontal component data enabled orientations of horizontal components to be relatively easily estimated onboard the scientific recovery cruise (Tsekhmistrenko et al., 2025).

2.2 OBS data noise

Compared to land stations, free fall OBSs are typically less sheltered from environmental conditions and their data are thus usually noisier. OBS-specific noise sources can be divided into four main categories: seafloor compliance, tilt noise, instrument-sediment coupling and current-induced vibrations (Duennebier et al., 1981; Lewis and Tuthill, 1981; Trehu, 1985a; Sutton and Duennebier, 1987; Crawford et al., 1991; Crawford and Webb, 2000; Stähler et al., 2018; Corela et al., 2023, among others). Tilt and seafloor compliance noise are both low frequency (<0.1 Hz) signals. The former is caused by the torque around the horizontal axis applied by ocean currents that make the instrument tilt periodically (Crawford et al., 1991; Webb, 1998). The latter is caused by the deformation of the seafloor by the passage of long-period ocean waves (Crawford and Webb, 2000). Instrument-sediment coupling noise is site-specific (Trehu, 1985b) and is present during the entire deployment as a narrowband signal within the 2 Hz to 15 Hz band.

Current-induced noise is generated by vortex shedding from the instrument frame, as well as auxiliary components such as antennas, flags, ropes, and floats. As water flows past the instrument, it generates vortices that apply lift and drag forces that cause the instrument to shake. The flow speed v, the characteristic dimension of the object L (usually its diameter), and the Strouhal number St (a dimensionless number that describes oscillating flow dynamics, Triantafyllou et al., 2016) affect the Strouhal frequency f_{vort}, or the frequency at which vortices are shed:

$$f_{vort} = St\frac{v}{L} \quad . \tag{1}$$

These vibrations are transmitted to the seismic sensor, where they are recorded as high-frequency signals above 1 Hz (Corela, 2014; Stähler et al., 2018; Essing et al., 2021; Corela et al., 2023). In some conditions, the Strouhal frequency approaches the natural oscillation frequency of the different OBS components, inducing a resonant state (Corela et al., 2023). Due to constructive interference, the amplitude of the vibrations increases suddenly as the lock-in regime is reached (Skop and Griffin, 1975; Griffin, 1985). In this regime, harmonics become discernible. The number of detectable harmonics may be affected by turbulent topographic wakes generated by seamounts and spillways (Tarakanov et al., 2018; Mashayek et al., 2024). Current-induced noise, or resonance noise, is instrument type-specific, with the same frequencies and number of harmonics appearing at multiple stations of the same type, but also site-specific, as seafloor topography can dramatically affect current velocities.

3 Methodology

3.1 Data processing and annotation

We computed 15 min spectrograms between 1 Hz and 50 Hz for vertical component data across all stations from the UPFLOW array. The precise details of the spectrogram computation and processing are given in Text S1 in the Supplementary Materials. The spectrograms were then linearly interpolated onto a target time-frequency grid. This had a time resolution of 15 s and a frequency resolution of 0.125 Hz, leading to 2D image-like data with dimensions $(60, 399)$ pixels in time-frequency. The choice of 15 min time windows (and the

fine frequency resolution) was motivated primarily by the characteristics of the resonance signals: they are highly narrowband and can persist for tens of minutes to hours, in contrast to more transient seismic events or whale vocalisations (e.g., Corela et al., 2023). Using longer windows ensures that these resonances are adequately captured while still maintaining a manageable data volume for manual annotation, training, and evaluation. These design choices were therefore made to prioritise robust representation of the resonance signal class.

We then manually annotated three datasets using the `Label Studio` package (Tkachenko et al., 2020-2024). The first was a training dataset using randomly sampled 15 min spectrograms from UPFLOW station UP05. We generated a dataset of 500 spectrograms and manually annotated five distinct feature classes: instrument resonances, blue whale vocalisations, earthquakes, ship noise, and an "other" category used for potentially interesting but unidentified signals. Three operators manually annotated the spectrograms, cross-checking work to ensure relatively consistent labelling criteria. The whole process took around 40 annotator-hours, though annotating sped-up considerably as we became more familiar with the signal types and annotation software. A pair of examples of manually annotated spectrograms is shown in Fig. 1. Of the classes considered here, resonances were by far the most common (both in pixel coverage and in number of objects), while blue whale vocalisations occurred significantly less frequently. The full class distribution is shown in Figure S1.

The second and third datasets were annotated for use as validation sets. We manually annotated instrument resonances in 50 spectrograms from the UPFLOW station UP34, as well as another 50 spectrograms from station RR40 of the OBS RHUM-RUM deployment in the Indian ocean in 2012-13 (Stähler et al., 2016; Barruol et al., 2017). Each chosen station had a different OBS type: UP05 had a "LOBSTER", UP34 had a "NAMMU", and the RR40 instrument was a "LCPO2000" (Stähler et al., 2016). This diversity of OBS types and geographic location, which leads to different noise backgrounds and instrument resonances, was chosen to test the robustness of the ML models. This approach allows for a controlled comparison of how differences in instrument type (as well as both instrument and deployment setting) affect model performance. It is distinct from randomly sampling spectrograms from the entire array, which would make interpretation more challenging. Again, all spectrograms were collected from times randomly sampled across the deployments.

Finally, spectrograms were globally clipped between a suitable dynamic range, scaled between $[0, 1]$ and zero padded to dimensionality $(64, 416)$ pixels. This last step was performed because several of the ML models required standardised input dimensions. The spectograms were input as single channels into the ML models (i.e., as greyscale images).

3.2 Machine learning techniques

A key motivation for this work was to leverage modern ML techniques for efficient and accurate feature dectection, particularly in situations where annotated training data are scarce. To this end, we utilised two frameworks for improving model performance in the low data regime: semi-supervised learning and transfer learning.

Semi-supervised learning algorithms are a popular approach to improve model performance when there is a limited set of annotated data (for reviews, see e.g. Van Engelen and Hoos, 2020; Yang et al., 2022). The principle behind semi-supervised learning is to jointly train on both labelled and unlabelled data. There is a wide range of approaches for utilising unlabelled data, of which perhaps the most popular is consistency regularization. This approach feeds two views of an unlabelled image, with different distortions or noise, and trains the ML model to ensure a consistent output over the two views.

Among ML methods, deep convolutional neural networks (CNNs) have been particularly influential in seismology, where they are widely applied to tasks such as arrival-time picking and denoising (e.g., Mousavi et al., 2020; Zhu and Beroza, 2018; Zhu et al., 2019; Yu et al., 2019; Münchmeyer et al., 2022; Bornstein et al., 2024; Trappolini et al., 2024). A key strength of CNNs is their ability to learn hierarchical, transferable feature representations (Sharif Razavian et al., 2014; He et al., 2016). This has motivated the use of pre-trained "foundation models," which can be then adapted to new tasks with limited labelled data (Tajbakhsh et al., 2016; Yosinski et al., 2014; Kornblith et al., 2019; Dosovitskiy et al., 2021; Zhai et al., 2022; Liu et al., 2022). This strategy is known as transfer learning and has proven especially valuable in domains with small, expensive-to-acquire labelled datasets (e.g., Hoffmann et al., 2019; Jain et al., 2022; Mishra et al., 2022; Hu et al., 2022).

In this work, we make extensive use of deep CNN models. We use an encoder-decoder model approach, which is a popular approach for semantic segmentation tasks (Badrinarayanan et al., 2017; Chen et al., 2018). An overview of this architecture is given in Fig. 2. In this framework, the encoder is utilised to extract features from the spectrograms (passed in as single channel greyscale images), while the decoder then composes these extracted features to make the class predictions on each pixel value. This work utilises a Residual Network (ResNet)18 for the encoder, a popular deep CNN that is well-suited for transfer learning tasks (He et al., 2016). A convolutional UNet architecture is then used as the decoder, which combines information over increasingly large spatial scales to produce the final pixel-level classifications (Ronneberger et al., 2015). We implement the neural network architecture in `segmentation_models_pytorch` (Iakubovskii, 2019), a high-level Pytorch-based library containing a range of prebuilt models that are frequently used for semantic segmentation.

It is important to note that we train a separate model for each feature class. We found this approach outper-

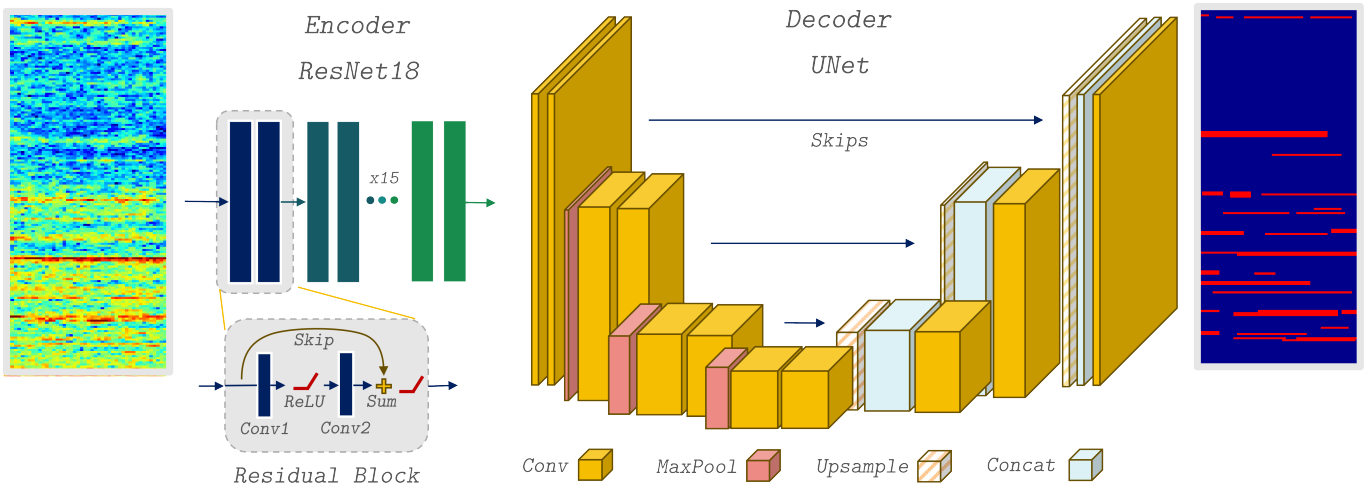

Figure 2 Semantic segmentation takes input data (spectrograms; left) and outputs a class for each pixel in the input (colored feature map; right). We use an encoder-decoder neural network architecture based on deep CNNs.

formed multiclass segmentation models. This slightly simplifies model training and evaluation. Each model outputs a probability mask of the same spatial dimensions as the input spectrogram, where each value between [0, 1] indicates the probability that a given pixel belongs to the target feature class. Pixels can then be assigned to the feature or background class by thresholding this probability.

3.3 Training Techniques

This section describes the three distinct training strategies we deployed to train the ML model:

- *Supervised learning*, which only uses manually annotated data as a baseline approach.
- *Semi-supervised learning*, which leverages unlabelled data through consistency regularisation via the Mean Teacher framework.
- *Synthetic transfer learning*, which pre-trains the model on a large set of synthetic examples before using the manually annotated data.

The latter approaches were chosen to mitigate the small size of the manually annotated dataset, which can significantly degrade the performance of deep ML models. We compare the performance of these approaches in Section 4. We performed a significant amount of heuristic-based hyperparameter optimisation, initially optimising architecture with the supervised learning approach, before optimising the training parameters (e.g., learning rates, training duration, batch sizes) for each approach separately. To provide a more reliable estimate of model performance, we repeated training several times with identical settings, with exact details provided below.

3.3.1 Supervised learning

The simplest approach to leverage our manually annotated dataset is to perform supervised learning; that is, to train the network to emulate the manual annotations made on the dataset. We follow standard practice by using a pixel-wise binary cross entropy loss $\mathcal{L}_{\text{sup}}$ on label masks $\mathbf{y}_{\text{label}} \in \{0, 1\}$ and predictions $\mathbf{y}_{\text{pred}}$:

$$\mathcal{L}_{\text{sup}}(\mathbf{y}_{\text{pred}}, \mathbf{y}_{\text{label}}) = \begin{aligned} &- [\mathbf{y}_{\text{label}} \cdot \log(\mathbf{y}_{\text{pred}}) \\ &+ w \cdot (1 - \mathbf{y}_{\text{label}}) \cdot \log(1 - \mathbf{y}_{\text{pred}})] \end{aligned} . \tag{2}$$

The null class feature weight w, which decreases the relative importance of the null class in the loss function, is a hyperparameter that can encourage the segmentation model to make more predictions of the target class (a common choice when dealing with severe class imbalances; Japkowicz and Stephen, 2002). This loss function trains the network to predict the probability that each pixel belongs to the target class. At inference time, a pixel is classified as belonging to the target class if $\mathbf{y}_{\text{pred}} > \tau$, where $\tau \in [0, 1]$ is the decision threshold. The default is typically $\tau = 0.5$, but τ can be tuned depending on the desired tradeoff between false positives and false negatives.

Note that even in the base supervised learning case, we initialise the ResNet encoder with weights from pre-training on `ImageNet` (Deng et al., 2009). This means we always perform a form of transfer learning, but in the supervised learning case we use a very generic pre-trained model that was only trained on natural images. This is in contrast with the highly specialised pre-training introduced in the synthetic data approach in Section 3.3.3.

3.3.2 Semi-supervised learning: The Mean Teacher framework

A key idea in semi-supervised learning is consistency regularisation, which posits that training a model to ensure consistent outputs under distorted views of the same data should improve robustness of the model (e.g., Laine and Aila, 2017; Sohn et al., 2020; Yang et al., 2022). This is particularly useful when dealing with very small annotated datasets, where consistency regularisation can be leveraged over the large unlabelled portion of a dataset (Fan et al., 2023). The mean teacher framework approaches this problem by tracking two models:

a student and a teacher model, and enforcing consistency between each of the model outputs (Tarvainen and Valpola, 2017). The student model is constantly trained via standard gradient descent on the loss function, while the teacher model weights are set to an exponential moving average of the student model weights. The semi-supervised loss function combines the supervised loss on labelled data with a consistency regularisation loss on unlabelled data:

$$\mathcal{L}_{\text{semi-sup}} = \mathcal{L}_{\text{sup}} + \lambda_{\text{cons}}\mathcal{L}_{\text{cons}} \quad , \tag{3}$$

where $\mathcal{L}_{\text{sup}}$ is identical to Eq. (2), $\mathcal{L}_{\text{cons}}$ is the consistency loss encouraging agreement between the teacher and student models, and λ_{cons} controls the weight of the consistency loss.

The consistency loss $\mathcal{L}_{\text{cons}}$ measures the disagreement between the student's predictions $\mathbf{y}_{\text{pred}}^{(s)}$ and the teacher's predictions $\mathbf{y}_{\text{pred}}^{(t)}$ on unlabelled data. For segmentation tasks, this can be implemented as a mean squared error between the logits:

$$\mathcal{L}_{\text{cons}} = \left\| \mathbf{y}_{\text{pred}}^{(s)} - \mathbf{y}_{\text{pred}}^{(t)} \right\|^2 \quad . \tag{4}$$

A crucial aspect of consistency regularisation is the distortion of input views through augmentations, the addition of noise, or both. This promotes robustness under realistic sources of noise, which can be particularly useful in improving the model's performance using unlabelled data. In this work, we designed a simple noise algorithm that injected global white noise, as well as transient broadband signals and persistent narrowband signals over random regions of the image. We randomly sampled around 5000 unlabelled spectrograms from the UP05 station for use as unlabelled data. These are demonstrated in Fig. 3. This noise is added to the student model data view, after which the semi-supervised loss function from Eq. (3) can be computed.

The teacher model's weights θ_t are updated every gradient descent step using an exponential moving average of the student model's weights θ_s:

$$\theta_t \leftarrow \alpha\theta_t + (1-\alpha)\theta_s \quad , \tag{5}$$

where α is the smoothing coefficient controlling how much the teacher depends on the current parameters of the student model. The teacher therefore provides a set of stable pseudo-labels for the unlabelled data that the student model is trained to match.

3.3.3 Synthetic data generation and transfer learning

We expect that generic natural image pre-training may be suboptimal for producing a useful feature extractor in the encoder given the very different characteristics of spectrograms. In addition, natural image datasets (e.g. Deng et al., 2009) tend to contain large objects in the image, so the resulting encoder may not be well suited to the fine-scale nature of both the instrument resonances and the whale calls we address in this work. Instead, by constructing a synthetic generator of the relevant features, the network can be pre-trained with a much more representative dataset. This allows the model to learn a better tailored set of generic feature extractors that are simpler to fine-tune on real data.

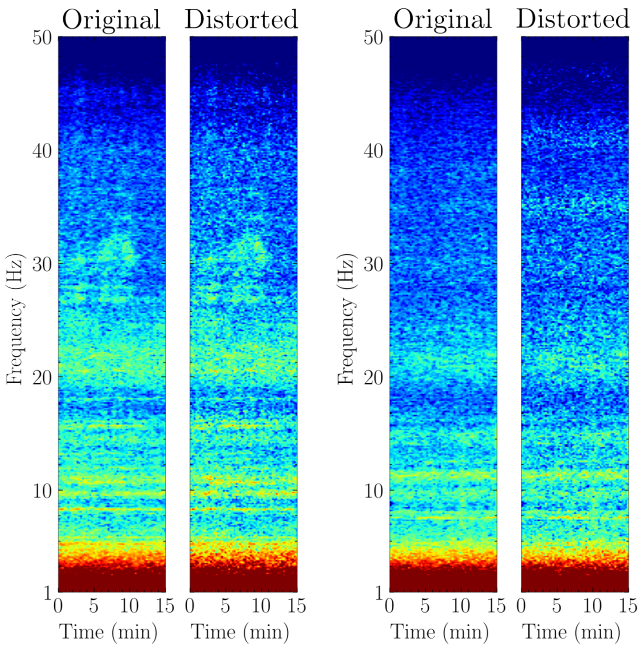

Figure 3 Two examples of pairs of original 15-min spectrograms and their slightly distorted counterpart. Minor white noise is added globally, as well as periodic persistent narrowband noise and (more subtle) transient wideband noise. These perturbations are used for consistency regularisation, as well as a data augmentation strategy in the synthetic pre-training scheme. The exact noise distortions can be found in the software (Saoulis et al., 2025).

For both the instrument resonance and the blue whale call classes, we created stochastic feature generators which produce realistic-looking features (see Fig. 4) according to a range of heuristics (e.g., number of features per image, rough location of the features, relationship between features). These were tuned to roughly match the appearance of the observed pattern of resonances and blue whale calls in the real data. A more detailed description of the synthetic feature generation procedure is provided in Text S2 in the Supplementary Materials, alongside examples of the synthetic feature templates (Figure S2) and a comparison between the synthetic and annotated feature characteristics (Figure S3).

The next step was to collect realistic background spectrograms that did not contain the relevant features, as training would otherwise be complicated by the existence of unlabelled features. We used a simple segmentation model trained in a supervised fashion for each of the target classes (following the procedure in Section 3.3.1), and used this to find spectrograms that *did not contain* any of the target class features. We found around 250 spectrograms with a very low probability of resonances, which were very rare. We visually inspected these spectrograms to validate that there was a very low rate of resonances. For the whale call detection task, we used a random set of 1600 spectrograms containing no blue whale calls (these were easy to find).

Finally, we combined the synthetic feature generators and background spectrograms with the noise distortions in Section 3.3.2. This addition of extra noise ensured further augmentation of the background spectrograms and synthetic features, allowing pre-training to continue for a large number of steps without overfitting. Synthetic pre-training was then performed by continuously sampling random background spectrograms, adding the stochastically generated synthetic features and noise, and then passing this mock data alongside the synthetic labels to the model for supervised training per Eq. (2). A visual demonstration of this procedure is provided in Figure S2. Examples of the resulting synthetic feature spectrograms and the corresponding known labels are given in Fig. 4. The exact implementation can be found in the software (Saoulis et al., 2025).

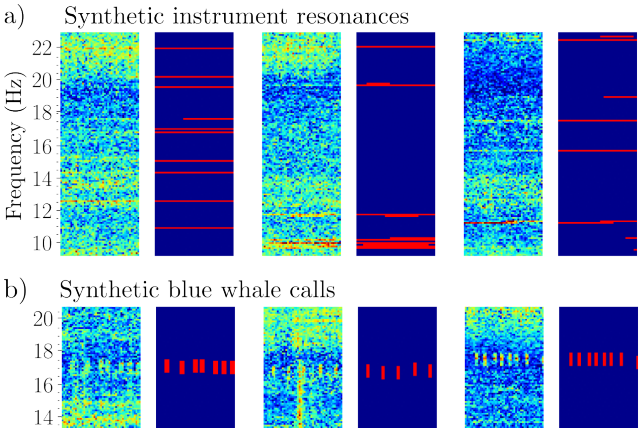

Figure 4 Three examples of synthetic feature generation in 15-min spectrograms and the associated labels for each of the feature types. a) shows the synthetic generation of instrument resonances, while b) shows the synthetic blue whale calls at 17 Hz.

3.4 Evaluation metrics

We rely on three metrics to evaluate model performance: precision, recall, and intersection over union (IoU). For false positives (FP), false negatives (FN), and true positives (TP), these metrics are defined as follows:

- Precision, which measures the fraction of the predicted positive pixels that are correct:

$$\text{Precision} = \frac{TP}{TP + FP} \quad . \tag{6}$$

- Recall, which measures the fraction of labelled positive pixels that are correctly predicted:

$$\text{Recall} = \frac{TP}{TP + FN} \quad . \tag{7}$$

- IoU, which measures the overlap between the predicted positive region and the ground truth positive region, defined as the ratio of their intersection to their union:

$$\text{IoU} = \frac{TP}{TP + FP + FN} \quad . \tag{8}$$

One common alternative metric used in binary classification tasks is the F1 score. Here, we instead rely on IoU as it is generally the headline metric in segmentation benchmarks, offers a clear geometrical interpretation, and is mathematically related to the F1 score, conveying the same information.

In this work, we found that the exact definition and region of each feature was often ambiguous (we see some evidence for this in our reported inter-annotator agreement statistics; see Table ST4). This led to small discrepancies across the dataset in terms of exactly how a feature was defined and annotated. In addition to this, some features were occasionally missed (particularly for the subtle, narrowband instrument resonances). We therefore argue that first IoU, and then recall, are more useful indicators for good model performance, as they do not overly penalize models for finding features that were missed during annotation. We also do not expect very high precision and recall scores, since the labels cannot be regarded as perfect "true" annotations.

4 Results

4.1 Model training

All training was performed on the 500 annotated UP05 spectrograms. We used an 80 %-20 % train - validation split throughout all experiments. Unless otherwise stated, we kept the validation set the same for all results to ensure comparable performance between runs. This left 400 spectrograms in the training set and 100 in the validation set. The two extra annotated datasets from UP34 and RR40, each containing 50 spectrograms, were not seen until all models had been optimised and trained.

Training was performed on a NVIDIA RTX A6000 GPU, with all training runs taking between $2 - 10$ minutes to run. For the supervised learning approach using annotated datasets, we trained resonance models for 20 epochs and whale call models for 60 epochs, where one epoch corresponds to a complete pass through of the entire manually annotated dataset. For the synthetic transfer learning approach, we fine-tuned the models by training on the annotated dataset for 8 epochs, finding that training for longer quickly led to overfitting. Since the concept of an epoch is less meaningful for the unsupervised and synthetic pre-training stages, we instead quote the total number of spectrograms processed during training below.

4.1.1 Hyperparameter selection

We systematically explored a broad range of architectures and hyperparameters. Key parameters were varied independently and the best-performing value was selected. This optimisation was performed manually between heuristically chosen ranges, rather than through exhaustive grid searches or automated procedures. The quantitative results of a number of representative experiments are provided in Table ST1 and Table ST2. Below, we summarise the optimal choices.

We evaluated multiple neural network architectures

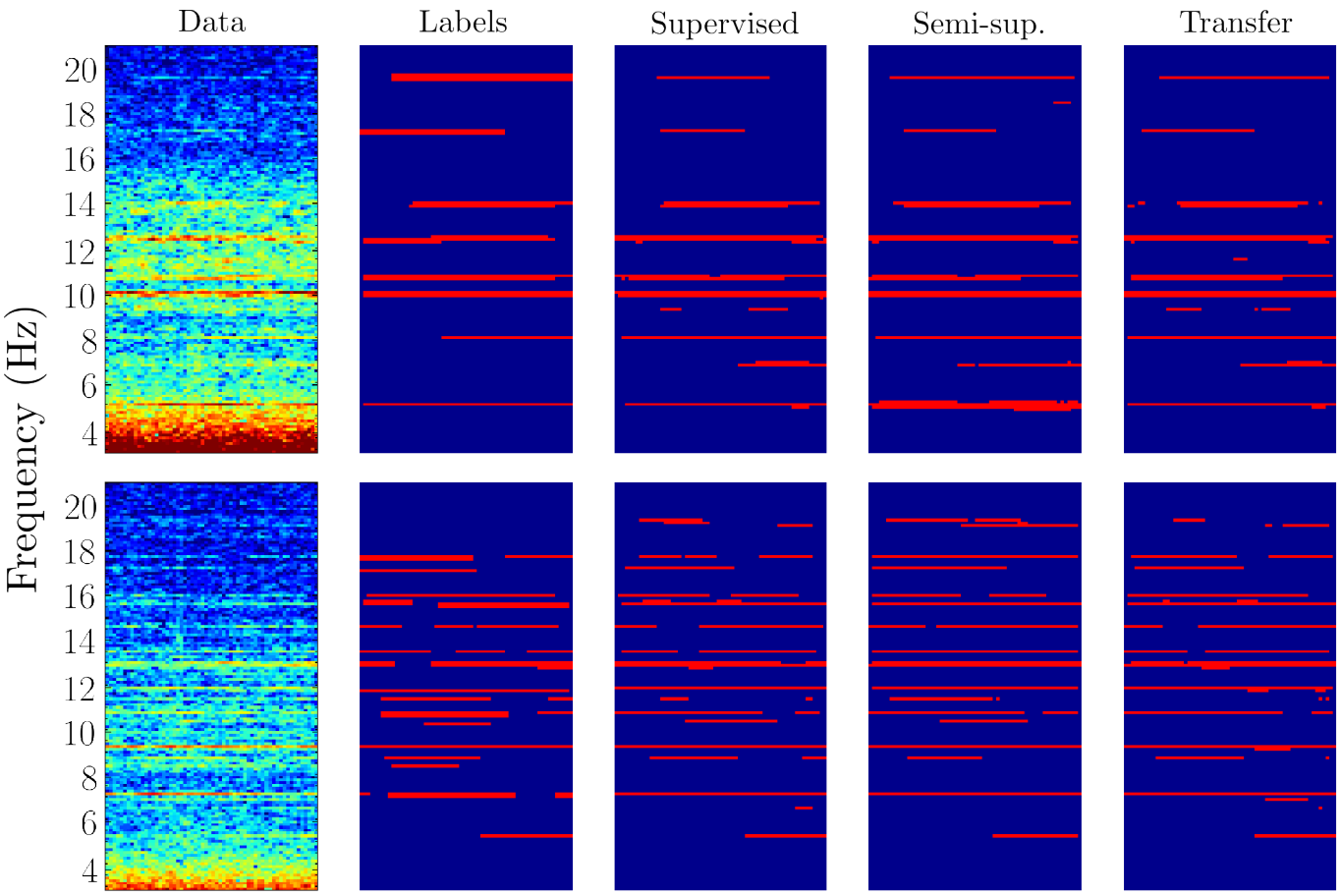

Figure 5 Two zoomed in examples of 15 min spectrograms with instrument resonances from the validation set, along with the associated manual labels. The three columns after show predictions from resonance segmentation models trained using each of the three approaches discussed in the main text. All three approaches show close agreement with the manual annotations, with some instances of incorrect predictions (e.g. top row, semi-supervised at 5 Hz) and cases where all models found resonances missed by the manual annotators (e.g. top row, 7 Hz).

for the encoder and decoder shown in Fig. 2. For the encoder, we tested Vision Transformers (ViTs; Dosovitskiy et al. 2021; Strudel et al. 2021), including MiT-b1 (Yu et al., 2023); for the decoder, we examined widely used segmentation architectures such as DeepLabV3+ (Chen et al., 2018) and Segformer (Xie et al., 2021). Although these models have been highly successful in many image segmentation applications, they proved unsuitable for our task as they operated at scales too large to annotate the fine-scale features in the spectrograms.

As such, we limited our search to several fully-convolutional networks. For the encoder, we tested various ResNets: ResNet18, ResNet50, and ResNet101, where the number denotes the number of residual convolutional blocks in the network. The shallowest network, ResNet18, performed as well as the deeper alternatives (this is consistent with previous passive acoustic whale detection work, see e.g. Bergler et al., 2019; Rasmussen and Širović, 2021), so we chose this network for its efficiency advantage. Performance across all tasks and training techniques was slightly improved by using ResNet18 weights that had been pre-trained on ImageNet. We did not see much change in performance when utilising more advanced decoders than the UNet, such as UNet++ (which includes a global attention mechanism at some stages of the network activations; Vaswani et al., 2017; Zhou et al., 2018). The performance of these architectures across both signal classes are reported in Table ST1. Finally, we found that training the network to perform binary segmentation always yielded better performance (a separate network for each task, rather than training a single network to perform multiclass prediction).

Once the architecture was settled, we performed hyperparameter optimisation for each training technique, with a representative set of results shown in Table ST2. We found that a small batch size of 10 spectrograms gave the best results (performing much better than larger batch sizes). We also utilised a weight-decay of 0.01 for all experiments, which regularises the neural networks by encouraging smaller magnitudes of the network weights (Krogh and Hertz, 1991), finding that it gave minor improvements. For the supervised objective $\mathcal{L}_{\text{sup}}$, we tuned the null class weight w: we set $w = 0.5$ for the resonance segmentation model and $w = 0.1$ for the whale segmentation model. The former choice led to improved performance (see Table ST2), while larger w values in the whale task led to unstable training, with the model ignoring all whale calls.

For the semi-supervised approach, we optimised the exponential moving average parameter for the teacher network, finding $\alpha = 0.98$ resulted in the best perfor-

Table 1 Metrics for each of the training techniques, evaluated on each of the evaluation datasets. The mean performance and standard error over three independent training runs are reported for each metric. Intersection over Union (IoU) for the UP05 validation set is presented to three decimal places as the standard error estimates were very small. In all cases, larger metric values correspond to better performance.

Dataset	UPFLOW UP05 Validation Set			UPFLOW UP34			RHUM-RUM RR40		
Metric	IoU	Precision	Recall	IoU	Precision	Recall	IoU	Precision	Recall
Supervised	0.420 ± 0.001	0.62 ± 0.03	0.57 ± 0.02	0.49 ± 0.02	0.56 ± 0.03	0.81 ± 0.02	0.52 ± 0.01	0.64 ± 0.02	0.73 ± 0.04
Semi-supervised	0.431 ± 0.002	0.62 ± 0.01	0.59 ± 0.01	0.47 ± 0.01	0.52 ± 0.01	0.85 ± 0.01	0.50 ± 0.02	0.63 ± 0.05	0.72 ± 0.06
Transfer	0.441 ± 0.002	0.63 ± 0.01	0.60 ± 0.01	0.53 ± 0.01	0.60 ± 0.01	0.83 ± 0.01	0.49 ± 0.01	0.63 ± 0.01	0.70 ± 0.02

mance. We used a linear ramp-up schedule that increased the consistency loss factor from $\lambda_{\text{cons}} = 0$ to $\lambda_{\text{cons}} = 0.5$ in Eq. (3) over the first 20 000 spectrograms passed into the model. To compute the semi-supervised objective function in Eq. (3) we used 10 labelled spectrograms for $\mathcal{L}_{\text{sup}}$ and 100 unlabelled spectrograms for the consistency loss contribution $\mathcal{L}_{\text{cons}}$ in every batch.

The main hyperparameters for the synthetic transfer learning were the training durations and learning rate of the two training stages, both of which use $\mathcal{L}_{\text{sup}}$. Pre-training on the synthetic data examples was performed for around 40 000 spectrograms. Fine-tuning using the manual annotations was then performed for 8 epochs with a low learning rate.

The exact configuration for all training runs can be found in the software (Saoulis et al., 2025).

4.2 Instrument resonance segmentation

In this section, we compare the results of each of the training approaches introduced above on the task of instrument resonance segmentation. This was by far the most frequently appearing feature in the dataset, with several hundred resonant pixels in the majority of the 500 annotated images (see Figure S1 for a comparison). We therefore expect that the semantic segmentation approach should have enough examples in the training dataset to ensure stable training and acceptable performance even without the improved training strategies. We set the resonance classification threshold to $\tau = 0.5$ for all results, finding that this gave the best IoU scores.

A comparison of segmentation results of the three methods is presented in Fig. 5. Broadly speaking, we see that all three models reproduce the majority of the human-annotated labelled resonances. There are still disagreements between the manual annotations and all models, and we validated that none of the models were capable of achieving the same level of performance as inter-annotator agreement (see Table ST4). However, the rate of false-positives is relatively low across all models. Careful inspection of Fig. 5 reveals that some of the ML models occasionally correctly identify some weak resonances that were missed by the human annotators.

A quantitative appraisal of each method is given in Table 1, where the model performance over various metrics was computed for the three datasets: the UP05 validation set, the UPFLOW UP34 evaluation set, and the RHUM-RUM RR40 evaluation set, as explained in Section 3.1. We repeated training using each method three times, with the mean and standard error for each metric given in the table. We validated that our evaluation dataset sizes were large enough to provide reliable model performance estimates by performing a bootstrapping analysis, with results presented in Table ST3.

As explained previously, we take IoU as a headline figure for performance, while precision and recall give a slightly more granular breakdown of the strengths and weaknesses of each model. The results show that the improved training strategies yield modest but significant improvements over the standard supervised learning strategy, between $0.01 - 0.02$ IoU (corresponding to around $2 - 5\%$). In addition, all models perform very well on the UP34 and RR40 datasets. This is particularly impressive since these are different instrument types to UP05, and therefore have different resonance characteristics than those seen in the training data. This is an indication of the strong generalisation performance of the models.

The results in Table 1 suggest a preference for a synthetic pre-training transfer learning strategy for the UPFLOW dataset. This is most clear from the IoU scores, where the synthetic pre-training approach performs 5% better than supervised learning on the UP05 validation dataset. We therefore utilise the best of these models for downstream applications on the UPFLOW array. However, the transfer learning approach performs slightly worse than the other two training strategies for the RHUM-RUM station (though with low statistical significance given the standard error estimates). One potential explanation could be the substantially different noise background and instrument resonance characteristics of the RHUM-RUM station. The long period of pre-training on UPFLOW-like noise and signals could have slightly overfit the model, leading to a minor drop in performance on this out-of-distribution data.

An account of the effects of hyperparameters is given in Table ST2, which indicates that hyperparameter improvements for the supervised approach could improve performance by a maximum of 0.02 – 0.03 IoU. In practice, the supervised model did not exceed ~ 0.42 IoU despite extensive architecture and hyperparameter experimentation. By contrast, synthetic pre-training achieved ~ 0.44 IoU, indicating that transfer learning provides performance gains beyond what was achievable through hyperparameter optimisation alone.

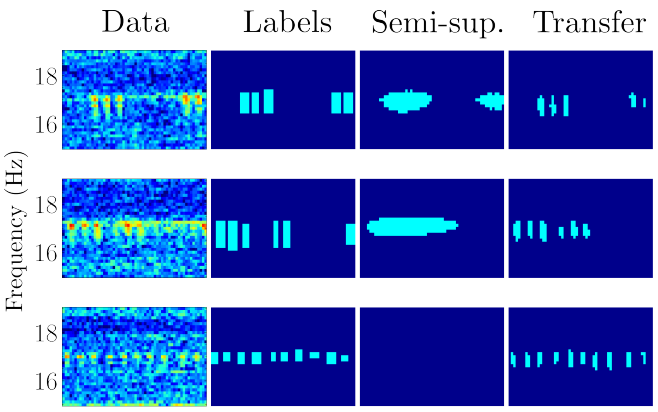

Figure 6 Left column: Three zoomed-in examples of 15-min spectrograms containing the 17 Hz blue whale call. The following columns show the associated manual labels, alongside predictions from models trained using semi-supervised learning and our synthetic transfer learning approach.

4.3 Blue whale call segmentation

We now explore the results on the blue whale call segmentation task. Blue whale calls could be observed at UP05 for several months during the deployment, though at intermittent times. As such, less than 10 % of spectrograms contained these whale calls, with a little over 200 individual vocalisations annotated in the UP05 dataset (significantly fewer than the resonance example; see Figure S1). These calls generally took the form of a periodic signal occurring at around 17 Hz. We found that using a greater probability threshold τ improved the model's ability to segment individual call regions, an important property for the applications discussed in Section 4.4.2. We therefore set the whale call classification threshold to $\tau = 0.8$ throughout.

We present qualitative examples of the semi-supervised and transfer learning approaches compared with the manual annotations in Fig. 6. In this example, there is a stark difference in the quality of the model predictions. The semi-supervised model is incapable of producing separate segmentation regions for each individual whale call. We find that this behaviour is replicated for various different hyperparameters, such as learning rate, class loss weighting, and architectures with both supervised and semi-supervised training. On the other hand, the synthetic transfer learning approach is capable of segmenting each individual whale call, with minor inaccuracies or differences with the manual annotations. This qualitative improvement is very valuable, given that the identification of individual calls is an essential precondition in order to associate calls and perform source location inversions.

Table 2 shows how each approach performs on the UP05 validation dataset. In order to produce reliable performance estimates given the small set of annotated whale calls, we perform repeat training multiple times over three different train-validation splits. The synthetic transfer learning approach significantly outperforms supervised and semi-supervised learning across all metrics, with an increase in the headline IoU metric from 0.100 ± 0.025 and 0.098 ± 0.031 respectively to 0.299 ± 0.027. We note that the uncertainties in the IoU performance are significantly greater than in Table 1. This is due to the very limited size of the annotated dataset, which leads to less stable training and greater variability in evaluation across such a small set of examples. Nonetheless, by considering the lower bound of the transfer learning result against the upper bounds of the weaker baselines we still find a (conservative) relative improvement in IoU of $\sim 100\%$. This is a very large improvement compared to the resonance segmentation discussed in Section 4.2. With a small dataset, supervised training was less stable and more susceptible to inconsistent labels, leading to degraded performance. On the other hand, pre-training on a large corpus of synthetic data enabled the model to first learn the correct structure of features, and it could thereafter be fine-tuned quickly and effectively on the small number of real examples.

Table 2 Blue whale call segmentation performance on the UP05 validation dataset by Intersection over Union (IoU), precision, and recall. The mean performance and standard error are reported across 9 runs, obtained from three distinct train/validation splits with three independent training repetitions per split.

Metric	IoU	Precision	Recall
Supervised	0.100 ± 0.025	0.21 ± 0.06	0.22 ± 0.07
Semi-supervised	0.098 ± 0.031	0.18 ± 0.05	0.22 ± 0.07
Transfer	0.299 ± 0.027	0.48 ± 0.05	0.49 ± 0.04

4.4 Applications

We ran inference using the best-performing transfer learning models for each signal class on the whole UPFLOW array. Spectrograms for the year-long UPFLOW deployment were computed in approximately 24 hours using 40 cores on a desktop class machine (around 2 s per spectrogram per core). Once array-wide spectrograms were computed, producing the prediction masks for 43 stations took around 2 h for each signal class on a desktop class machine with a NVIDIA RTX A6000 GPU. The runtime varied approximately linearly with spectrogram count (and therefore station count), meaning that a full year of data from a single OBS station could be processed in approximately 3 min. In addition, model inference could be run on the desktop CPUs at around 35 spectrograms per second. Our framework could therefore be deployed to run in real time on consumer-grade hardware.

4.4.1 Tidally driven instrument resonances

The movement of tidal water masses plays an important role in the modulation of permanent and seasonal ocean bottom currents because it can reinforce or subdue the prevailing flow patterns depending on its phase, amplitude, and interaction with local bathymetry (for e.g., van Geel et al., 2020; Bailey et al., 2024). Given that OBS resonances are in large part driven by water flowing past the instrument, we therefore expect

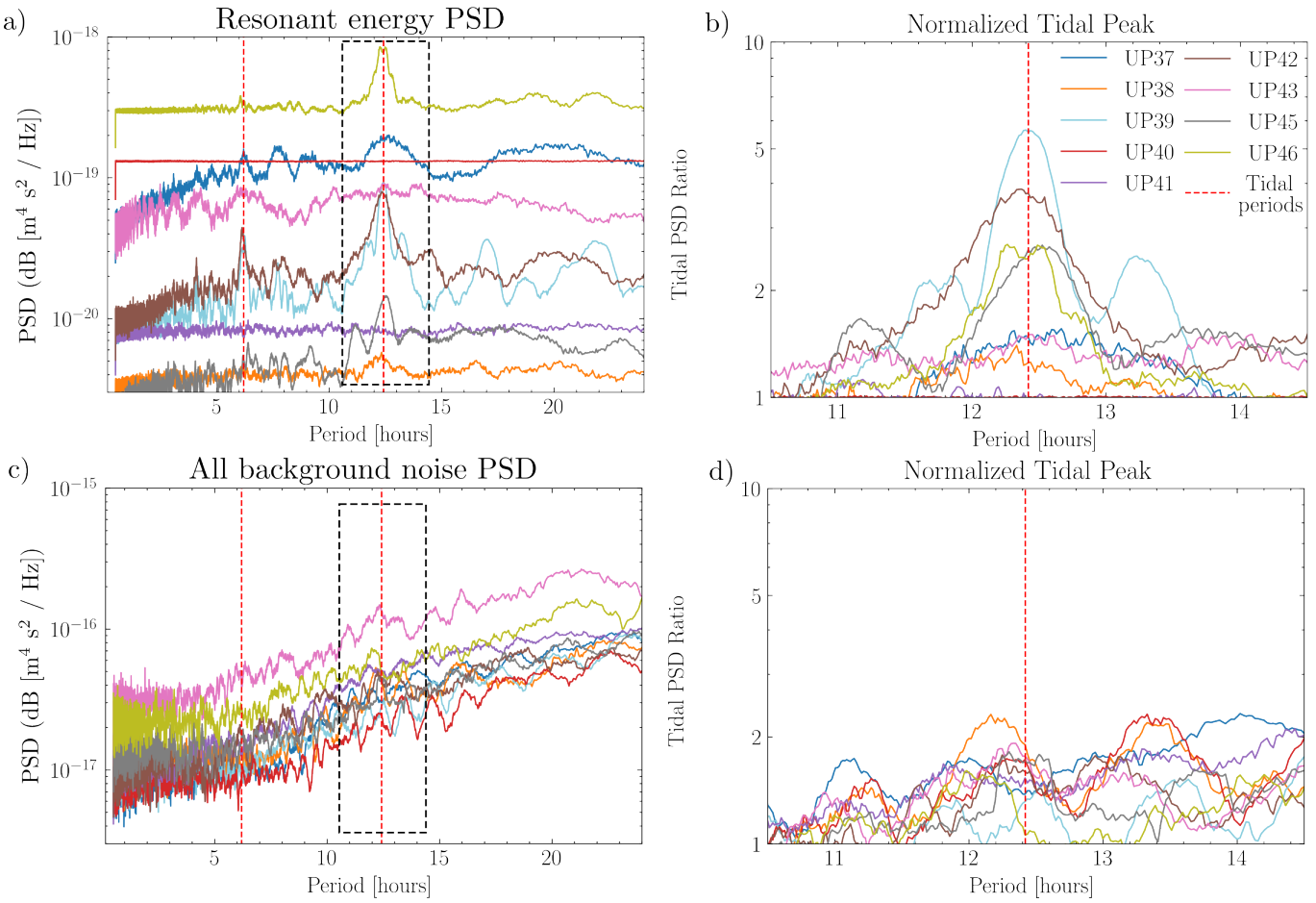

Figure 7 *Upper panels*: panel a) shows the resonant energy power spectral density (PSD) curves for a subset of the UPFLOW stations around Madeira Island (Fig. 1). Red dashed lines indicate the expected periods of tidal peaks at 12 h 25 mins and 6h 12 mins. A zoom in on the main tidal peak (dashed black box) is shown in b), where the ratio between the PSD peak amplitude and the surrounding background PSD is computed for each station. *Lower panels*: plots c) and d) repeat the analysis of a) and b) for the raw spectrogram data (i.e. unprocessed with no separation between resonances and other sources of noise).

to find some relationship between tidal-imprinted currents and the energy deposited in the instrument resonances.

In order to estimate the resonant energy, we used the segmentation model to produce masks identifying resonant and non-resonant regions in the spectrograms. Resonant energy was obtained by extracting the power of the resonant pixels and subtracting the background power spectrum. The mean background power spectrum was estimated from the non-resonant pixels by averaging across 0.75 Hz bands of the 15 min spectrogram. The width of this band was chosen to exceed the width of the largest resonant bands ($\sim$ 4 pixels, or 0.5 Hz), ensuring a reliable background estimate. This procedure yielded a year-long time-series of resonant energy for each station with a sampling rate of 15 s (the time-resolution of the spectrograms), which we reduced to a 15 min sampling rate by averaging over each spectrogram. For the analysis in the main text, we simply summed the resonant energy contributions across all frequency bands.

We first explored whether a tidal signal could be extracted from the resonant energy series. Fig. 7 shows the power spectral density (PSD) of these time series for a subset of UPFLOW stations surrounding Madeira Island (Fig. 1). For reference, Fig. 7 also presents the PSD of the raw spectrograms, without any distinction between resonances and background. Under idealised circumstances, the tide-induced currents have a frequency of double the tidal frequency (i.e. at a period of 6.25 h). However, the prevailing mid-Atlantic current roughly counteracts the tidal currents in one direction around Madeira, leading to a weaker first peak at 6.25 h and a stronger peak at double the period of 12.5 h (Corela et al., 2023). This effect is clearly visible in Fig. 7, and the exact degree of suppression of the first peak and the amplification of the second peak depends on the specific station. This is qualitatively expected, given that exposure to each type of current is highly dependent on station location and sea bottom spillways. Further work is required for more quantitative interpretations. Importantly, Fig. 7 shows that the extracted resonant energy has a much stronger tidal signature than the raw spectrograms.

In Text S6 in the Supplementary Materials, we demonstrate that our segmentation approach also enables a more granular, frequency-specific analysis. We find that a very narrow frequency band between 6.5 – 7.5 Hz contains a strong tidal signal, and again demonstrate that our resonance segmentation model signif-

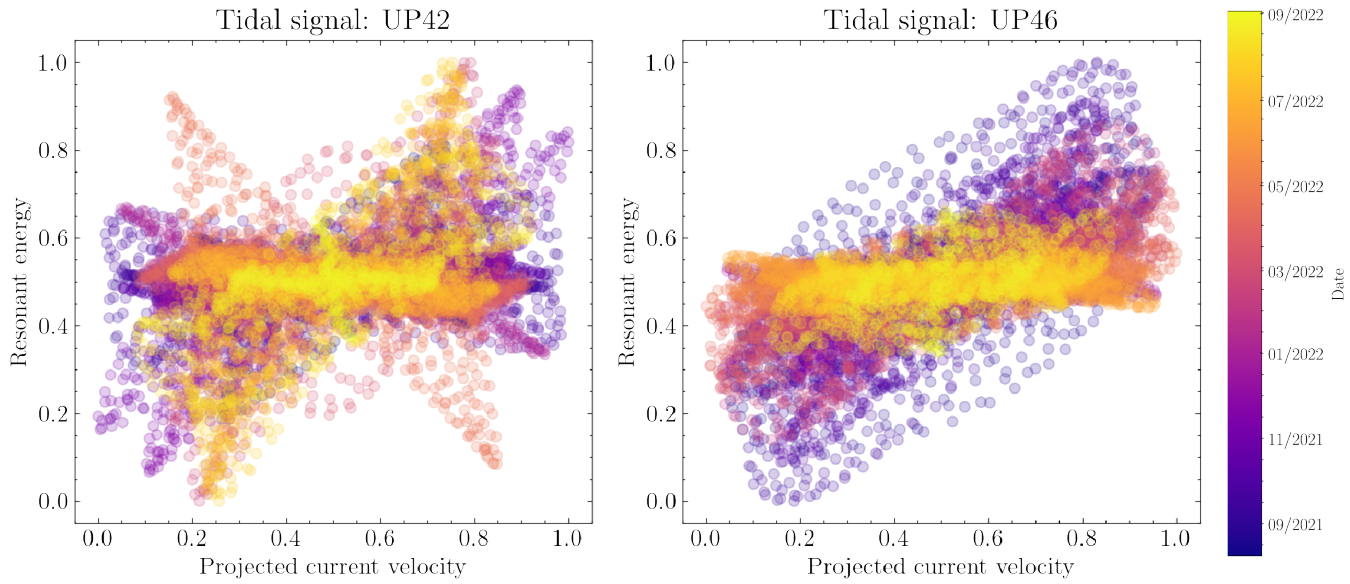

Figure 8 Scatter plots of the projected current velocity (see main text for details) at each station against the energy deposited in the instrument resonances for UPFLOW stations UP42 and UP46. Each raw time series has been bandpass filtered between 11 h and 13 h to extract the tidal signal, and normalised to between $[0, 1]$. The complex structure of the scatter plots indicates a non-linear, time-varying relationship between the current data product and the measured resonances. The raw time series for station UP46 are presented in Fig. S5.

icantly amplifies the tidal signal relative to the background noise. The existence of a single resonance excited by tidal currents suggests that instrument resonances recorded by OBSs have the potential to physically disambiguate between different current regimes.

We then more directly probed whether the resonant energy can be used as a proxy for current. To do this, we extracted a time-series of barotropic current velocities at each of the UPFLOW stations from the Copernicus Atlantic-Iberian Biscay Irish-Ocean Physics Reanalysis dataset (Copernicus Marine Service, 2024). Prior analysis has demonstrated that the strength of instrument resonances can strongly depend on the directions of currents (Corela et al., 2023). We therefore constructed a projected velocity quantity that related the ocean current to the OBS instrument orientation, with details given in Text S3 in the Supplementary Materials.

Both the projected current and resonant energy time series were bandpass-filtered between periods of 11 h and 13 h for each station. Both time series were then normalised between $[0, 1]$ on a per-station basis. A scatter plot of the relationship between the resulting resonant energy and current time series is shown for UP42 and UP46 in Fig. 8.

We find that a complex relationship between tidal currents and resonant energy emerges for these two stations. This relationship is time-varying, with different lobes in the scatter plot activating over different periods of the deployment. The raw, unfiltered time series of these quantities is presented in Figure S5, which demonstrates periods of strong correlation. While the currents continually oscillate, strong periods of resonances only activate under certain conditions. The existence of this non-random, time-varying structure within the scatter plots indicates some latent relationship between the two variables. Uncertainty in the reliability of the current product at the seafloor complicates interpretation, especially given the likely impacts of local effects such as seafloor topography. For other stations, particularly when the tidal energy peak in Fig. 7 is not as strong, there is no clear relationship in the scatter plot.

This example demonstrates a methodological goal of this study, i.e. accurate and precise feature detection using semantic segmentation. However, building a predictive model of ocean current from OBS instrument resonances is not achievable in this study. Substantial additional work is needed to calibrate the mapping between resonances and currents (notably through the collection of current-meter data), and to establish the robustness of this approach across different stations and instrument types.

4.4.2 Blue whale call detection across the UPFLOW array

We repeated the spectrogram segmentation procedure from Section 4.4.1 with the best-performing blue whale call segmentation model.

To restrict detections to the blue whale call region and reduce false positives, we applied a simple post-processing step: the predicted whale call mask was multiplied with a binary frequency mask spanning 16–18 Hz, thereby retaining only detections within the 17 Hz blue whale band. This simple frequency masking step illustrates a unique strength of the spectrogram segmentation approach: prior expert knowledge can be incorporated directly into the analysis at the prediction stage.

Once the whale call prediction masks were generated for the whole array, we processed these into call counts by finding the number of connected prediction regions containing more than 6 pixels. This threshold

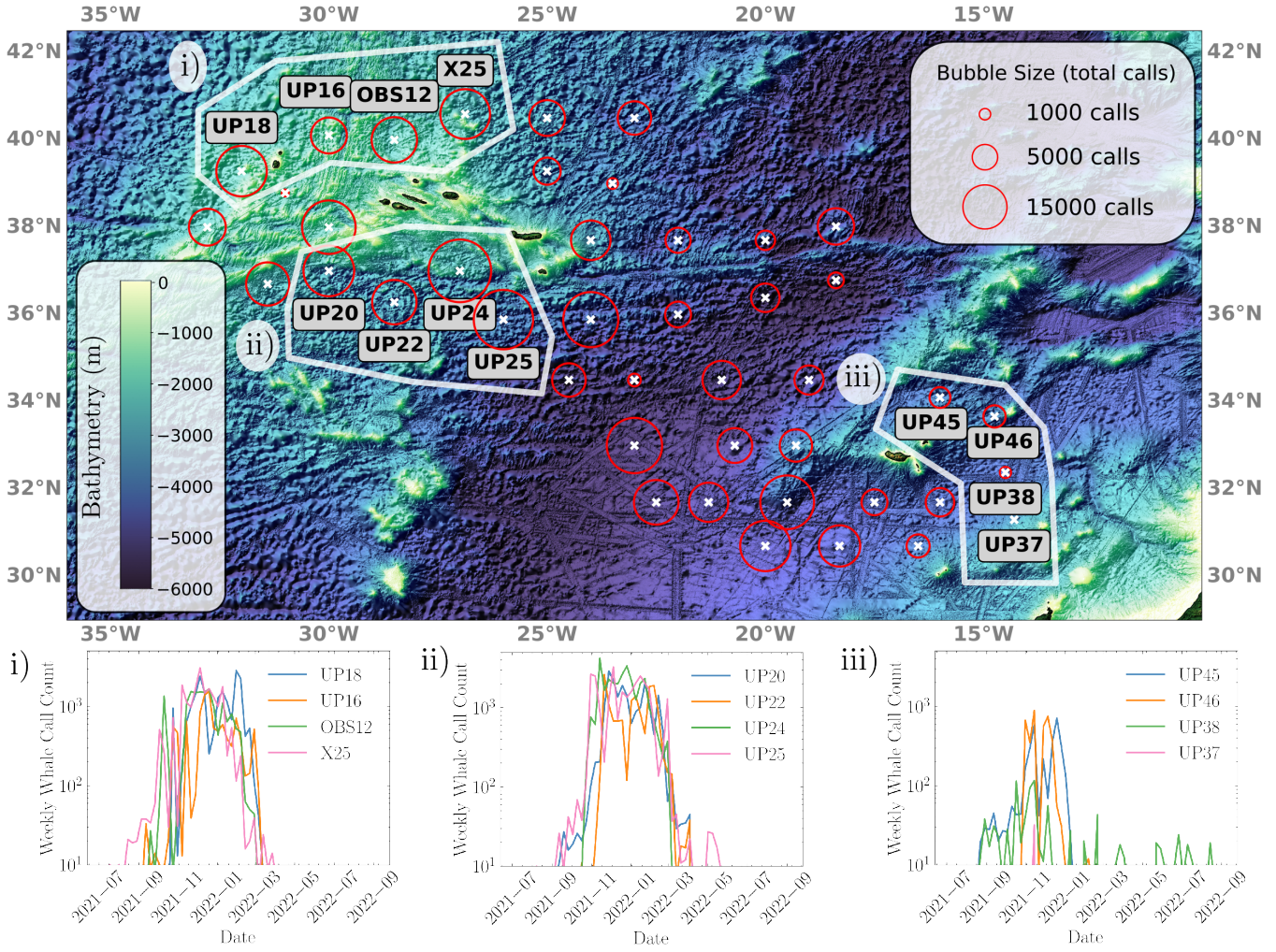

Figure 9 Blue whale call counts across the entire deployment. Model predictions produced by the synthetic transfer learning model. Main: the total number of counts at each station, visualised by the bubble size at each station. Insets: temporal breakdowns of the total call count for a few station subsets. An animation showing the weekly whale call counts at each station is provided in the Supplementary Materials (see Figure S4).

provided a satisfactory balance between reducing false positives and identifying clear, high SNR whale vocalisations. Note that individual whale call segmentation, and therefore call counting, was only possible using the model trained via transfer learning (Fig. 6).

Fig. 9 shows the spatial and temporal distribution of the detected whale counts across the UPFLOW array. We find that blue whale calls are detected across the entire array, with over 1000 calls detected at 39 of the 43 OBSs analysed. Only at UP37, in the far south-east of the array, were less than 40 calls detected, several of which were false positives. Fig. 9 shows significant spatial variations in the whale call detections: many of the stations north and south of the Azores recorded a very large number of whale calls (> 10000), while towards the north-east and east of the array there are significantly fewer calls recorded at most stations (< 5000). A full table of detected call counts is given in Table ST6. We find the seasonal distribution of the whale calls closely matches previous work (Romagosa et al., 2020), with the vast majority of detected calls occurring between mid-October to mid-March.

An animation of the spatial distribution of whale calls by week of the deployment is available in the Supplementary Materials (Figure S4). This animation shows that the first arrival of blue whale calls appears around August, concentrated to the north of the Azores islands. Over the next three months, we see significant whale call activity across the entire deployment, matching the general pattern in Fig. 9. Finally, towards the end of the season in February, we see a migration of calls towards the southern side of the OBS deployment, which gradually tapers away. Our results are consistent with the expected seasonal distribution of whale calls in the Azores islands (Romagosa et al., 2020); however, it is unclear whether the observed spatio-temporal variations in the blue whale vocalisations are caused by genuine changes in the location distribution of the whales, or by the well-documented seasonality of the 17 Hz call, which is not produced in the spring and summer months (Lockyer, 1984; Stafford et al., 2001; Akamatsu et al., 2014; Romagosa et al., 2020).

We also performed a quantitative analysis of the blue whale vocalisation detection false positive rate. The migration patterns and seasonal behaviour of the blue whales also provide a proxy for the false positive rate.

Over the summer months, no blue whale vocalisations are expected in the Azores region (Romagosa et al., 2020). We analysed this period and found a detection rate, and therefore false positive rate, of 4.4 calls per station per week. In order to further validate this, we manually assessed model detections over around 200 hours of data sampled uniformly across the array, detailed in Text S5 in the Supplementary Materials. We found that our manually verified out-of-season false positive rate of 5.9 ± 2.5 was consistent with the 4.4 figure. However, during the active season the false positive rate was substantially higher (15.6 ± 5.3 per station per week), particularly during periods of intense calling activity when earthquakes and other broadband transients were sometimes misclassified as calls. Even so, these false positives represent only $\sim 1-3\%$ of the total detections, indicating that the vast majority of detections remain reliable. The full details and further interpretation of this analysis are presented in Text S5 in the Supplementary Materials. We also present a range of examples of automated call detections in Figure S7, demonstrating the relatively low false positive rate.

We investigated the spikes in false positive rates at UP25 and UP38 outside the vocalisation season of blue whales (shown in Fig. 9 ii) and iii) respectively). In both cases, we found that the false positive rate was driven by an onset of a repeatedly occurring impulsive signal. At UP25, this lasted for a few weeks, while at UP38 the signal recurred for the whole second half of the deployment. An example of these signals is shown in Figure S7. The physical origin of these short-duration events is unclear, and further work would be required to determine whether they match the wide range of signals previously reported in the literature (such as gas-related processes at the seafloor Batsi et al., 2019). Nevertheless, the rarity of these false positives highlights the reliability of the ML-based blue whale call detection algorithm as applied to the UPFLOW array.

Interestingly, we detect many blue whale calls around Madeira, where blue whale sightings are very rare (Fernandez et al., 2021; Valente et al., 2019). Indeed, Freitas et al. (2012) reported a record of only 4 blue whale sightings in Madeiran waters, as of 2012. While we manually verified that the vast majority of these detections were genuine blue whale calls, the distance of the whales to the OBSs is not clear (see Figure S7 for a range of examples). There is a wide range of maximum detection distances for low frequency passive acoustic surveys in the literature, ranging from $\mathcal{O}(10)$ km (e.g., Harris et al., 2013; Hilmo and Wilcock, 2024) to >100 km (e.g., Širović et al., 2007; Samaran et al., 2010; Kuna and Nábělek, 2021; Wilcock and Hilmo, 2021). Recent work has shown that blue whale calls can be detected at ranges of up to 1000 km (De Castro et al., 2024), with strong directionality preferences induced by factors including bathymetry and seafloor properties. These factors, combined with the relatively low signal-to-noise ratio of many of the detections at the Madeira stations, indicate further analysis is required to confirm whether these calls are being made near Madeira.

4.4.3 Whale tracking through source localisation

Finally, we present a proof-of-concept exploration of whether the synthetic transfer learning blue whale detection model can be applied to whale tracking. This exercise is intended to illustrate potential applications, rather than to provide a robust localisation analysis.

There is a wide range of strategies that can be leveraged for whale call source location (e.g., Harris et al., 2013; Dréo et al., 2019; Pereira et al., 2020; Hilmo and Wilcock, 2024). Advanced localisation techniques, such as multipath ranging approaches and three-component analysis, can in principle improve accuracy, but they require both detailed propagation models (bathymetry, sediment properties, sound-speed profiles) and the ability to resolve precise arrival times and multipath structure in the recordings. However, our coarse temporal resolutions (15 s) cannot resolve these features at the required precision, meaning that higher-fidelity propagation models would be unlikely to yield meaningful improvement. We therefore adopt standard arrival-time localisation as a pragmatic choice.

Unfortunately, the inter-station distance of the UPFLOW array is large enough to make call association challenging, since inter-station arrival time differences approach and surpass typical inter-call intervals. Instead, we applied our model to an example from the RHUM-RUM array, where inter-station distances between some stations are much shorter. We took an Antarctic blue whale tracking example from the array studied by Dréo et al. (2019), where whale calls were manually picked and localisations were performed with a multi-path ranging technique. We used our model to generate whale call predictions across the same subset of stations used in Dréo et al. (2019). Given the different call frequency of the Antarctic blue whale call, for this application we multiplied the model predictions with a different mask centered on the expected 25 Hz whale vocalization band. This allowed us to generate a catalogue of timestamped calls for all five stations surrounding the whale track.

We then associated call arrivals that occurred within an interval 30 s of each other. For calls with at least 3 associated arrivals, we utilised the equal differential time (EDT) likelihood (Lomax et al., 2000), a 2-D arrival time difference algorithm. In this framework, the parameter σ represents the assumed arrival-time uncertainty entering the EDT likelihood. We adopted a conservative value of $\sigma = 30$ s to absorb small timing errors in the segmentation-derived call timestamps. We assumed a constant sound velocity of 1.5 km s^{-1} everywhere. This produced a 2D probability density field for each associated call. Due to the very low time resolution, each source location solution had very broad uncertainties. To account for this, we stacked all source location solutions in 30 minute intervals to produce aggregated probability density fields over time. The results of this analysis are presented in Fig. 10, overlaid with the results from Dréo et al. (2019). Note that accurate whale vocalisation timestamps were only possible due to the pixel-level segmentation predictions of the model.

We find good qualitative agreement between the two

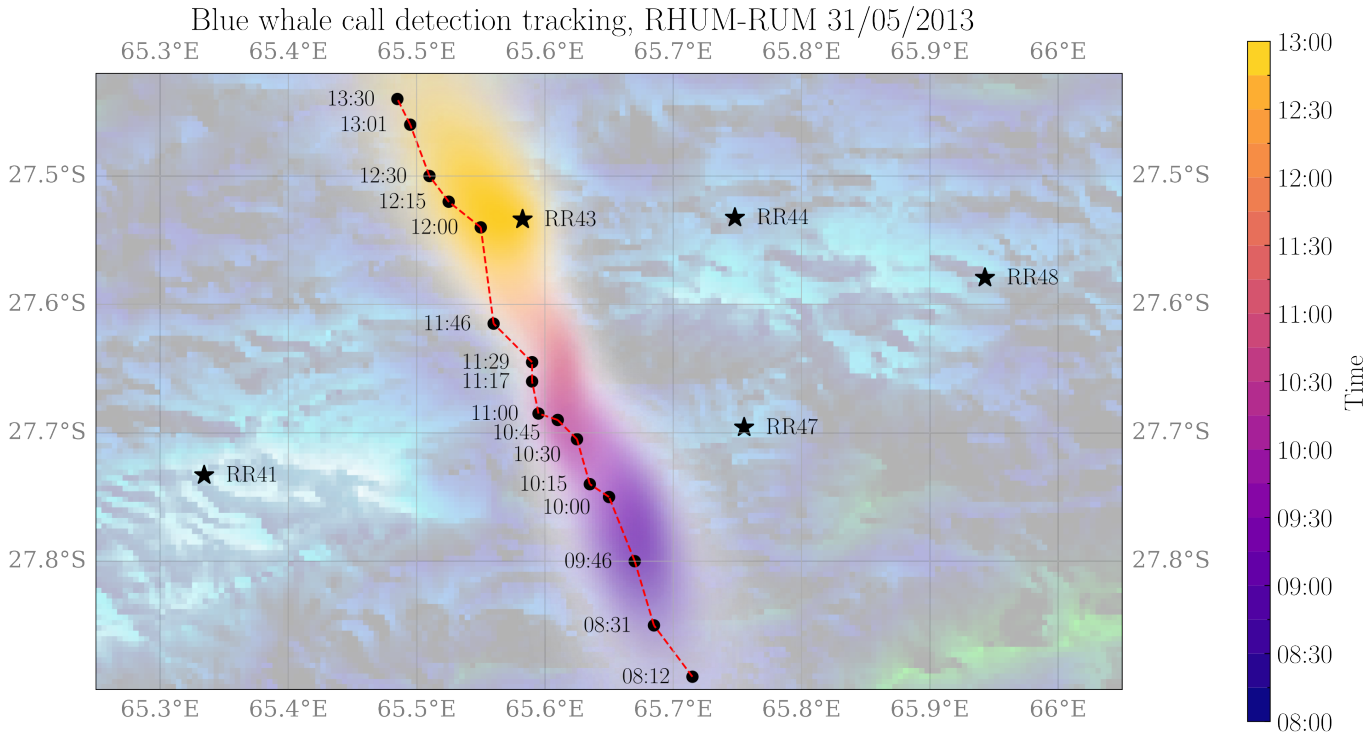

Figure 10 Comparison between our automated probabilistic tracking algorithm (coloured by time) for blue whale calls recorded by RHUM-RUM stations RR41, RR43, RR44, RR47 and RR48 in the Indian ocean on 31 May 2013, overlaid with the tracking results from Dréo et al. (2019). The latter were produced by manual arrival time picking (shown by the black markers and the dashed red line).

tracks, indicating that our whale call detection model is capable of producing call timestamps that are sufficiently accurate for exploratory localisation analyses. At the same time, the limitations of this simple approach are evident in the inferred track. The location resolution becomes very poor at the start and end of the track, potentially due to the weak constraining power of the station distribution. There is also an eastward bias of our track. We hypothesise that this may be caused by our simplistic propagation model, which ignores bathymetry and ocean layering that can affect the potential detection of different ray paths at each station. These oversimplifications may also be exacerbated by the asymmetric station distribution.

5 Discussion

This study introduces semantic segmentation for time-frequency representations of seismic data. We have demonstrated the feasibility of semantic segmentation to quickly and accurately identify a range of non-seismic features in OBS data. We showed that a small manually annotated dataset is sufficient to fine-tune deep learning models that have been pre-trained on natural image data. Moreover, we were able to significantly improve model performance by designing a simple synthetic pre-training dataset for each signal type, outperforming the more generic semi-supervised learning framework. Our ML-based feature detection approach was capable of generalising from simple, simulated features to real data with only a small set of realistic examples, demonstrating its flexibility.

We then validated our methodology across a range of examples: tracking the relationship between instrument resonances and tidally-driven ocean currents, detecting blue whale vocalisations across the UPFLOW array, and recovering a blue whale track from the RHUM-RUM experiment (Dréo et al., 2019). This was all achieved after training on just 500 (randomly selected) spectrograms from a single OBS station. The fact that performance remained strong across instruments and deployments highlights the robustness of our approach, even without explicit cross-domain adaptation. While domain adaptation strategies (e.g., per-station fine-tuning, normalization, and latent representation alignment) could further improve transferability, our results suggest that acceptable performance can already be achieved in a zero-shot setting.

We also demonstrated that our approach enables accurate detection of features while requiring several orders of magnitude fewer annotations than traditional deep learning pickers. The method could in principle achieve even greater label efficiency by focusing annotation efforts on specific target signals. Although these results are demonstrated here for resonances and whale calls, similar strategies could in principle be applied to other seismic phenomena with distinctive time–frequency signatures. Future work could explore whether the techniques introduced here could serve as a starting point for detecting and analysing signals such as volcanic tremors, short-duration events, earthquake foreshocks, or deep earthquakes, where datasets may be limited or site-specific.

Our pixel-level annotation of spectrograms is unique

among prior work in seismic signal detection. It provides more fine-grained, interpretable predictions than previous work performing segment classification or 1-D time series segmentation. Time-frequency segmentation enables automated extraction of detailed feature characteristics, such as energy deposited in resonances and whale call timestamps, as demonstrated here, with a wide range of possible future applications (e.g., whale vocalisation duration and frequency). In addition, our time-frequency representation approach places this work in contrast to the widespread use of time-series representations in seismology. This draws on the dominant trend in audio processing, which has largely shifted toward time-frequency representations as input for deep learning models (e.g., Dieleman and Schrauwen, 2014; Van den Oord et al., 2013; Piczak, 2015; Miller et al., 2023; Stowell, 2022). The data efficiency and high accuracy observed here is likely related to this well-chosen data representation, as convolutional neural networks (CNNs) excel at learning the localised spectral-temporal patterns of desired features. This can improve generalisation by enhancing discrimination of overlapping signals and noise, particularly at high frequencies.

5.1 Limitations and future directions

Our methodology has some limitations. Manual annotation, particularly the pixel-level annotation required for semantic segmentation, is relatively time-consuming. To this end, more advanced ML techniques such as weakly-supervised learning, which learns segmentation regions from image-level class labels, could be explored (Wei et al., 2016; Ahn and Kwak, 2018; Zhou, 2017). Moreover, where the desired model output is the detection of discrete events, such as whale vocalisations and earthquakes, the object detection framework may be a more desirable alternative to semantic segmentation (e.g., Redmon et al., 2016; He et al., 2017; Wang et al., 2024; see Zou et al., 2023; Terven et al., 2023 for reviews). Additionally, while CNNs are translationally equivariant and therefore robust to local time-shifts, our approach does not capture continuity across windows. This can limit robustness at window edges. This could be improved by leveraging sliding-window consistency checks, or at least by applying the model over overlapping windows.

The design of our pre-training datasets was driven by very simple heuristics (e.g., the rough incidence rate, frequency, and amplitude of the features). Note that our heuristic-driven synthetic data generation was made simple by the native time-frequency representation. Improved pre-training techniques could involve more realistic signal and noise simulations, such as observation-driven whale call emulation. In addition, Table 1 showed that synthetic pre-training slightly degraded performance when transferring from UPFLOW data to RHUM-RUM data, potentially as a result of a mismatch between the background noise profiles. Future work could explore whether generalisability to different stations could be improved by incorporating unlabelled data from more stations in the training phase (i.e. directly in the semi-supervised approach, or by using background noise spectrograms from many stations in the synthetic pre-training stage).

While we found that binary class segmentation models performed best for supervised learning, multi-class segmentation would be significantly more computationally efficient. It is also possible that, with larger annotated datasets, multi-class training could mitigate against some of the systematic false positives identified in this work, such as earthquakes being misidentified as whale calls. To this end, future work could explore whether synthetic pre-training could be leveraged for multiple signal classes at once to enable acceptable performance in a multi-class model. We leave exploration of the other feature classes in our annotated dataset to future work.

There remains a wide range of challenges that must be addressed before our methodology can be applied in a more systematic way. While we show instrument resonances can be treated as a proxy for currents at the seafloor, current-meter data collected close to the OBS (e.g. replicating Godin et al., 2024; Tan et al., 2025 for OBSs, or in the lab as in Wu et al., 2023), are required to build calibrated models of the current. Such models could potentially allow for current estimations across both UPFLOW and historical OBS deployments, at least for a given instrument type. For blue whale density estimation, quantitative estimates of the spatio-temporal false positive rates are required (see e.g., Marques et al., 2013; De Castro et al., 2024). In addition, extending the methodology to other marine mammal species (e.g. fin whales), as well as producing more accurate localisations and characterisations of the calls, requires higher time-frequency resolution, as well as a more accurate, expert-labelled dataset.

Finally, it would be desirable to perform a comparison between the methodology presented here and alternative techniques for identifying instrument noise (e.g., Essing et al., 2021; Zali et al., 2023) and marine mammal detections (e.g., Allen et al., 2021; Plourde and Nedimović, 2022; Stowell, 2022; Napoli and White, 2023). This could guide future efforts to improve the training techniques and architectures used to identify both common and uncommon features in OBS data. In bioacoustics, a few earlier traditional approaches also performed pixel-level classification (e.g., Roch et al., 2011; Gillespie et al., 2013), but Jin et al. (2022) appears to be the only prior study to apply semantic segmentation to whale vocalisations, as we have done here (Rasmussen and Širović, 2021; Cotillard et al., 2024 instead use object detection frameworks). Most other previous studies use the more standard feature extraction or ML-based binary segment classification frameworks.

6 Conclusion

In conclusion, this study presents a flexible, data-efficient approach for detecting a wide range of features in OBS data, such as instrument reverberations and blue whale calls, using semantic segmentation for time-frequency representations of seismic data. We demonstrated that with specialised training techniques,

accurate feature detection can be achieved with relatively few annotations, all while maintaining robustness across instruments and deployments. Our pixel-level time-frequency framework also enables fine-grained, interpretable analyses that go beyond traditional time-series or segment-level classification, unlocking a wide range of potential applications.

Acknowledgements

We thank Carlos Corela for providing insight on ocean currents in the mid-Atlantic, and its impact on OBSs. We also thank Andreia Pereira, Miriam Romagosa, Mónica Silva, and Gabrielle Arrieta for valuable discussions in interpreting the blue whale vocalisation detection data. We thank Vaibhav Vijay Ingale and an anonymous reviewer for their constructive feedback during the review stage. We are grateful to all the institutions that contributed instruments to the UPFLOW experiment: 32 OBSs were rented from the DEPAS international pool of instruments maintained by the Alfred Wegener Institute (Bremerhaven), Germany, while additional instruments were borrowed from other institutions (7 from DIAS, 4 from IDL, 3 from ROA, and 4 from GEOMAR).

Data and code availability

This study utilised UPFLOW data, which are deposited in the GFZ EIDA node (Ferreira, 2024) (network 8J, data embargoed until May 2028). The RHUM-RUM data are publicly available via RESIF (Barruol et al., 2017). The current data used is available through the E.U. Copernicus Marine Service Information (Copernicus Marine Service, 2024). Code to produce the spectrogram data, run training and inference of the ML models, and the manual UP05 annotations dataset are available on GitHub: `https://github.com/asaoulis/reverb`, as well as via the Zenodo repository Saoulis et al. (2025). The study used `PyTorch` version `2.0.0` throughout, with key library versions given in the GitHub repository Saoulis et al. (2025). This work made extensive use of the `Python` libraries `Label-Studio` (Tkachenko et al., 2020-2024) and `segmentation_models_pytorch` (Iakubovskii, 2019).

Competing interests

The authors have no competing interests, nor do they work for, advise, own shares in, or receive funds from any organisation that could benefit from this article. They have declared no affiliations apart from their research organisations.

Funding

We thank the UPFLOW project, funded by the European Research Council under the European Union's Horizon 2020 research and innovation program (grant agreement No 101001601). AAS was supported by the STFC UCL Centre for Doctoral Training in Data Intensive Science (grant ST/W00674X/1) and by departmental and industry contributions. AAS was also supported by the A. G. Leventis Foundation educational grant scheme. AL was funded by the Portuguese Fundação para a Ciência e a Tecnologia (FCT) I.P./MCTES through national funds (PIDDAC) – UIDB/50019/2020; UIDP/50019/2020; LA/P/068/2020.

References

Ahn, J. and Kwak, S. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4981–4990, 2018.

Akamatsu, T., Rasmussen, M. H., and Iversen, M. Acoustically invisible feeding blue whales in Northern Icelandic waters. *The Journal of the Acoustical Society of America*, 136(2):939–944, 2014. doi: 10.1121/1.4887439.

Allen, A. N., Harvey, M., Harrell, L., Jansen, A., Merkens, K. P., Wall, C. C., Cattiau, J., and Oleson, E. M. A convolutional neural network for automated detection of humpback whale song in a diverse, long-term passive acoustic dataset. *Frontiers in Marine Science*, 8:607321, 2021. doi: 10.3389/fmars.2021.607321.

An, C., Cai, C., Zhou, L., and Yang, T. Characteristics of low-frequency horizontal noise of ocean-bottom seismic data. *Seismological Society of America*, 93(1):257–267, 2022. doi: 10.1785/0220200349.

Anthony, R. E., Aster, R. C., Wiens, D., Nyblade, A., Anandakrishnan, S., Huerta, A., Winberry, J. P., Wilson, T., and Rowe, C. The seismic noise environment of Antarctica. *Seismological Research Letters*, 86(LA-UR-14-28568), 2014. doi: 10.1785/0220140109.

Aster, R. C., McNamara, D. E., and Bromirski, P. D. Multidecadal climate-induced variability in microseisms. *Seismological Research Letters*, 79(2):194–202, 2008. doi: 10.1785/gssrl.79.2.194.

Badrinarayanan, V., Kendall, A., and Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017. doi: 10.1109/TPAMI.2016.2644615.

Bailey, L. P., Clare, M. A., Hunt, J. E., Kane, I. A., Miramontes, E., Fonnesu, M., Argiolas, R., Malgesini, G., and Wallerand, R. Highly variable deep-sea currents over tidal and seasonal timescales. *Nature Geoscience*, July 2024. doi: 10.1038/s41561-024-01494-2.

Barruol, G., Davy, C., Fontaine, F. R., Schlindwein, V., and Sigloch, K. Monitoring austral and cyclonic swells in the "Iles Eparses"(Mozambique channel) from microseismic noise. *Acta Oecologica*, 72:120–128, 2015. doi: 10.1016/j.actao.2015.10.015.

Barruol, G., Sigloch, K., RHUM-RUM Group, and RESIF. RHUM-RUM experiment, 2011-2015, code YV (Réunion Hotspot and Upper Mantle – Réunion's Unterer Mantel) funded by ANR, DFG, CNRS-INSU, IPEV, TAAF, instrumented by DEPAS, INSU-OBS, AWI and the Universities of Muenster, Bonn, La Réunion, 2017. https://seismology.resif.fr/networks/#/YV__2011. doi: 10.15778/RESIF.YV2011.

Batsi, E., Tsang-Hin-Sun, E., Klingelhoefer, F., Bayrakci, G., Chang, E. T., Lin, J.-Y., Dellong, D., Monteil, C., and Géli, L. Nonseismic signals in the ocean: Indicators of deep sea and seafloor processes on ocean-bottom seismometer data. *Geochemistry, Geophysics, Geosystems*, 20(8):3882–3900, 2019. doi: 10.1029/2019GC008349.

Baumgartner, M. F. and Mussoline, S. E. A generalized baleen whale call detection and classification system. *The Journal of*

the Acoustical Society of America, 129(5):2889–2902, 2011. doi: 10.1121/1.3562166.

Bell, S. W., Forsyth, D. W., and Ruan, Y. Removing noise from the vertical component records of ocean-bottom seismometers: Results from year one of the Cascadia Initiative. *Bulletin of the Seismological Society of America*, 105(1):300–313, 2015. doi: 10.1785/0120140054.

Bergler, C., Schröter, H., Cheng, R. X., Barth, V., Weber, M., Nöth, E., Hofer, H., and Maier, A. ORCA-SPOT: An automatic killer whale sound detection toolkit using deep learning. *Scientific reports*, 9(1):10997, 2019. doi: 10.1038/s41598-019-47335-w.

Bornstein, T., Lange, D., Münchmeyer, J., Woollam, J., Rietbrock, A., Barcheck, G., Grevemeyer, I., and Tilmann, F. PickBlue: Seismic phase picking for ocean bottom seismometers with deep learning. *Earth and Space Science*, 11(1):e2023EA003332, 2024. doi: 10.1029/2023EA003332.

Brodie, D. C. and Dunn, R. A. Low frequency baleen whale calls detected on ocean-bottom seismometers in the Lau basin, southwest Pacific Ocean. *The Journal of the Acoustical Society of America*, 137(1):53–62, 2015. doi: 10.1121/1.4904556.

Bromirski, P. D., Duennebier, F. K., and Stephen, R. A. Mid-ocean microseisms. *Geochemistry, Geophysics, Geosystems*, 6(4), 2005. doi: 10.1029/2004GC000768.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. Encoder-decoder with Atrous Separable Convolution for Semantic Image Segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, September 2018.

Chen, Y., Ji, D., Ma, Q., Zhai, C., and Ma, Y. A Novel Generative Adversarial Network for the Removal of Noise and Baseline Drift in Seismic Signals. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024. doi: 10.1109/TGRS.2024.3358901.

Choi, S., Lee, B., Kim, J., and Jung, H. Deep-learning-based seismic-signal p-wave first-arrival picking detection using spectrogram images. *Electronics*, 13(1):229, 2024. doi: 10.3390/electronics13010229.

Copernicus Marine Service. Atlantic-Iberian Biscay Irish-Ocean Physics Reanalysis, 2024. doi: 10.48670/moi-00029. IBI_MULTIYEAR_PHY_005_002. Ocean physical reanalysis product for the Iberia-Biscay-Ireland (IBI) region, provided by the IBI Monitoring and Forecasting Centre (IBI-MFC). Accessed on 6 March 2025.

Corela, C. *Ocean bottom seismic noise : applications for the crust knowledge, interaction ocean-atmosphere and instrumental behaviour*. Phd thesis, Faculdade de Ciências da Universidade de Lisboa, 2014. http://hdl.handle.net/10451/15805.

Corela, C., Loureiro, A., Duarte, J. L., Matias, L., Rebelo, T., and Bartolomeu, T. The effect of deep ocean currents on ocean-bottom seismometers records. *Natural Hazards and Earth System Sciences*, 23(4):1433–1451, 2023. doi: 10.5194/nhess-23-1433-2023.

Cotillard, T., Sécheresse, X., Aubin, J., Mikus, M.-A., Vergara, V., Gambs, S., Michaud, R., Martins, C. C., Turgeon, S., Chion, C., et al. Automatic detection and classification of beluga whale calls in the St. Lawrence estuary. *The Journal of the Acoustical Society of America*, 156(6):3723–3740, 2024. doi: 10.1121/10.0030472.

Crawford, W. C. and Webb, S. C. Identifying and Removing Tilt Noise from Low-Frequency (<0.1 Hz) Seafloor Vertical Seismic Data. *Bulletin of the Seismological Society of America*, 90(4): 952–963, aug 2000. doi: 10.1785/0119990121.

Crawford, W. C., Webb, S. C., and Hildebrand, J. A. Seafloor compliance observed by long-period pressure and displacement measurements. *Journal of Geophysical Research: Solid Earth*, 96 (B10):16151–16160, September 1991. doi: 10.1029/91jb01577.

Dahmen, N. L., Clinton, J. F., Meier, M.-A., Stähler, S. C., Ceylan, S., Kim, D., Stott, A. E., and Giardini, D. MarsQuakeNet: A more complete marsquake catalog obtained by deep learning techniques. *Journal of Geophysical Research: Planets*, 127(11): e2022JE007503, 2022. doi: 10.1029/2022JE007503.

Davy, C., Barruol, G., Fontaine, F. R., Sigloch, K., and Stutzmann, E. Tracking major storms from microseismic and hydroacoustic observations on the seafloor. *Geophysical Research Letters*, 41 (24):8825–8831, 2014. doi: 10.1002/2014GL062319.

De Castro, F. R., Harris, D. V., Buchan, S. J., Balcazar, N., and Miller, B. S. Beyond counting calls: estimating detection probability for Antarctic blue whales reveals biological trends in seasonal calling. *Frontiers in Marine Science*, Volume 11:1406678, 2024. doi: 10.3389/fmars.2024.1406678.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. doi: 10.1109/CVPR.2009.5206848.

Dieleman, S. and Schrauwen, B. End-to-end learning for music audio. In *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 6964–6968. IEEE, 2014. doi: 10.1109/ICASSP.2014.6854950.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

Dréo, R., Bouffaut, L., Leroy, E., Barruol, G., and Samaran, F. Baleen whale distribution and seasonal occurrence revealed by an ocean bottom seismometer network in the Western Indian Ocean. *Deep Sea Research Part II: Topical Studies in Oceanography*, 161:132–144, 2019. doi: https://doi.org/10.1016/j.dsr2.2018.04.005.

Duennebier, F. K., Blackinton, G., and Sutton, G. H. Current-generated noise recorded on ocean bottom seismometers. *Marine Geophysical Researches*, 5(1):109–115, 1981. doi: 10.1007/bf00310316.

Dunn, R. A. and Hernandez, O. Tracking blue whales in the eastern tropical Pacific with an ocean-bottom seismometer and hydrophone array. *The Journal of the Acoustical Society of America*, 126(3):1084–1094, 2009. doi: 10.1121/1.3158929.

Essing, D., Schlindwein, V., Schmidt-Aursch, M. C., Hadziioannou, C., and Stähler, S. C. Characteristics of Current-Induced Harmonic Tremor Signals in Ocean-Bottom Seismometer Records. *Seismological Research Letters*, apr 2021. doi: 10.1785/0220200397.

Fan, Y., Kukleva, A., Dai, D., and Schiele, B. Revisiting consistency regularization for semi-supervised learning. *International Journal of Computer Vision*, 131(3):626–643, 2023. doi: 10.1007/s11263-022-01723-4.

Fernandez, M., Alves, F., Ferreira, R., Fischer, J.-C., Thake, P., Nunes, N., Caldeira, R., and Dinis, A. Modeling fine-scale cetaceans' distributions in oceanic islands: Madeira Archipelago as a case study. *Frontiers in Marine Science*, 8: 688248, 2021. doi: 10.3389/fmars.2021.688248.

Ferreira, A. M. G. Upward mantle flow from novel seismic observations (UPFLOW), 2024. Other/Seismic Network [Dataset]. Available at: https://geofon.gfz-potsdam.de/waveform/archive/network.php?ncode=8J&year=2021.

Freitas, L., Dinis, A., Nicolau, C., Ribeiro, C., and Alves, F. New records of cetacean species for Madeira Archipelago with an updated checklist. *Boletim do Museu Municipal do Funchal (História Natural)*, 334(LXII):25–43, 2012.

Gaspà Rebull, O., Cusí, J. D., Ruiz Fernández, M., and Muset, J. G. Tracking fin whale calls offshore the Galicia Margin, north east Atlantic Ocean. *The Journal of the Acoustical Society of America*, 120(4):2077–2085, 2006. doi: 10.1121/1.2336751.

Gillespie, D., Caillat, M., Gordon, J., and White, P. Automatic detection and classification of odontocete whistles. *The Journal of the Acoustical Society of America*, 134(3):2427–2437, 09 2013. doi: 10.1121/1.4816555.

Godin, O. A., Tan, T. W., Joseph, J. E., and Walters, M. W. Observation of exceptionally strong near-bottom flows over the Atlantis II Seamounts in the northwest Atlantic. *Scientific Reports*, 14(1): 10308, 2024. doi: 10.1038/s41598-024-60528-2.

Goodwin, M., Halvorsen, K. T., Jiao, L., Knausgård, K. M., Martin, A. H., Moyano, M., Oomen, R. A., Rasmussen, J. H., Sørdalen, T. K., and Thorbjørnsen, S. H. Unlocking the potential of deep learning for marine ecology: overview, applications, and outlook. *ICES Journal of Marine Science*, 79(2):319–336, 01 2022. doi: 10.1093/icesjms/fsab255.

Griffin, O. M. *Vortex-induced vibrations of marine cables and structures*. Naval Research Laboratory, 1985.

Gualtieri, L., Camargo, S. J., Pascale, S., Pons, F. M., and Ekström, G. The persistent signature of tropical cyclones in ambient seismic noise. *Earth and Planetary Science Letters*, 484:287–294, 2018. doi: 10.1016/j.epsl.2017.12.026.

Harris, D., Matias, L., Thomas, L., Harwood, J., and Geissler, W. H. Applying distance sampling to fin whale calls recorded by single seismic instruments in the northeast Atlantic. *The Journal of the Acoustical Society of America*, 134(5):3522–3535, 2013. doi: 10.1121/1.4821207.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

Hilmo, R. and Wilcock, W. S. D. Estimating distances to baleen whales using multipath arrivals recorded by individual seafloor seismometers at full ocean depth. *The Journal of the Acoustical Society of America*, 155(2):930–951, 02 2024. doi: 10.1121/10.0024615.

Hoffmann, J., Bar-Sinai, Y., Lee, L. M., Andrejevic, J., Mishra, S., Rubinstein, S. M., and Rycroft, C. H. Machine learning in a data-limited regime: Augmenting experiments with synthetic data uncovers order in crumpled sheets. *Science advances*, 5(4): eaau6792, 2019. doi: 10.1126/sciadv.aau6792.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Iakubovskii, P. Segmentation Models Pytorch. https://github.com/qubvel/segmentation_models.pytorch, 2019.

Jain, S., Seth, G., Paruthi, A., Soni, U., and Kumar, G. Synthetic data augmentation for surface defect detection and classification using deep learning. *Journal of Intelligent Manufacturing*, pages 1–14, 2022. doi: 10.1007/s10845-020-01710-x.

Jansson, A., Humphrey, E., Montecchio, N., Bittner, R., Kumar, A., and Weyde, T. Singing voice separation with deep U-Net convolutional networks. In *18th International Society for Music Information Retrieval Conference*, pages 23–27, 2017.

Japkowicz, N. and Stephen, S. The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5):429–449, 2002. doi: 10.3233/IDA-2002-650.

Jiang, C., Fang, L., Fan, L., and Li, B. Comparison of the earthquake detection abilities of PhaseNet and EQTransformer with the Yangbi and Maduo earthquakes. *Earthquake Science*, 34(5): 425–435, 2021. doi: 10.29382/eqs-2021-0038.

Jin, C., Kim, M., Jang, S., and Paeng, D.-G. Semantic segmentation-based whistle extraction of Indo-Pacific Bottlenose Dolphin residing at the coast of Jeju island. *Ecological Indicators*, 137: 108792, 2022. doi: 10.1016/j.ecolind.2022.108792.

Kedar, S., Longuet-Higgins, M., Webb, F., Graham, N., Clayton, R., and Jones, C. The origin of deep ocean microseisms in the North Atlantic Ocean. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 464(2091):777–793, 2008. doi: 10.1098/rspa.2007.0277.

Koper, K. D. and Burlacu, R. The fine structure of double-frequency microseisms recorded by seismometers in North America. *Journal of Geophysical Research: Solid Earth*, 120(3):1677–1691, 2015. doi: 10.1002/2014JB011820.

Koper, K. D., Burlacu, R., Armstrong, A. D., and Herrmann, R. B. Classifying small earthquakes, explosions and collapses in the western United States using physics-based features and machine learning. *Geophysical Journal International*, 239(2): 1257–1270, 2024. doi: 10.1093/gji/ggae316.

Kornblith, S., Shlens, J., and Le, Q. V. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2661–2671, 2019.

Krogh, A. and Hertz, J. A simple weight decay can improve generalization. *Advances in neural information processing systems*, 4, 1991.

Kuna, V. M. and Nábělek, J. L. Seismic crustal imaging using fin whale songs. *Science*, 371(6530):731–735, feb 2021. doi: 10.1126/science.abf3962.

Laine, S. and Aila, T. Temporal Ensembling for Semi-Supervised Learning. In *International Conference on Learning Representations*, 2017.

LA/P/068/2020, 2020. doi: 10.54499/LA/P/0068/2020.

Lapins, S., Goitom, B., Kendall, J.-M., Werner, M. J., Cashman, K. V., and Hammond, J. O. A little data goes a long way: Automating seismic phase arrival picking at Nabro volcano with transfer learning. *Journal of Geophysical Research: Solid Earth*, 126(7): e2021JB021910, 2021. doi: 10.1029/2021JB021910.

Lewis, B. T. R. and Tuthill, J. D. Instrumental waveform distortion on ocean bottom seismometers. *Marine Geophysical Researches*, 5(1):79–85, 1981. doi: 10.1007/bf00310313.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022.

Lockyer, C. Review of baleen whale (Mysticeti) reproduction and implications for management. *Reports of the International Whaling Commission*, 6:27–50, 1984.

Lomax, A., Virieux, J., Volant, P., and Berge-Thierry, C. Probabilistic earthquake location in 3D and layered models: Introduction of a Metropolis-Gibbs method and comparison with linear locations. *Advances in seismic event location*, pages 101–134, 2000. doi: 10.1007/978-94-015-9536-0$_5$.

Marques, T. A., Thomas, L., Martin, S. W., Mellinger, D. K., Ward, J. A., Moretti, D. J., Harris, D., and Tyack, P. L. Estimating animal population density using passive acoustics. *Biological reviews*, 88 (2):287–309, 2013. doi: 10.1111/brv.12001.

Mashayek, A., Gula, J., Baker, L. E., Naveira Garabato, A. C., Cimoli, L., Riley, J. J., and de Lavergne, C. On the role of seamounts in upwelling deep-ocean waters through turbulent mixing. *Proceedings of the National Academy of Sciences*, 121(27), June 2024. doi: 10.1073/pnas.2322163121.

Matias, L. and Harris, D. A single-station method for the detection, classification and location of fin whale calls using ocean-bottom seismic stations. *The Journal of the Acoustical Society of America*, 138(1):504–520, 2015. doi: 10.1121/1.4922706.

McDonald, M. A., Hildebrand, J. A., and Webb, S. C. Blue and fin whales observed on a seafloor array in the Northeast Pacific. *The Journal of the Acoustical Society of America*, 98(2):712–721, 1995. doi: 10.1121/1.413565.

Miller, B. S., 15, I.-S. A. T. W. G. M. B. S. . S. K. M. . V. O. I. . H. D. . S. F. . Š. A. . B. S. . F. K. ., Balcazar, N., Nieukirk, S., Leroy, E. C., Aulich, M., Shabangu, F. W., Dziak, R. P., Lee, W. S., and Hong, J. K. An open access dataset for developing automated detectors of Antarctic baleen whale sounds and performance evaluation of two commonly used detectors. *Scientific Reports*, 11(1):806, 2021a. doi: 10.1038/s41598-020-78995-8.

Miller, B. S., Calderan, S., Leaper, R., Miller, E. J., Širović, A., Stafford, K. M., Bell, E., and Double, M. C. Source level of Antarctic blue and fin whale sounds recorded on sonobuoys deployed in the deep-ocean off Antarctica. *Frontiers in Marine Science*, 8: 792651, 2021b. doi: 10.3389/fmars.2021.792651.

Miller, B. S., Madhusudhana, S., Aulich, M. G., and Kelly, N. Deep learning algorithm outperforms experienced human observer at detection of blue whale D-calls: a double-observer analysis. *Remote Sensing in Ecology and Conservation*, 9(1):104–116, 2023. doi: 10.1002/rse2.297.

Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., and Terzopoulos, D. Image segmentation using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3523–3542, 2021. doi: 10.1109/TPAMI.2021.3059968.

Mishra, S., Panda, R., Phoo, C. P., Chen, C.-F. R., Karlinsky, L., Saenko, K., Saligrama, V., and Feris, R. S. Task2sim: Towards effective pre-training and transfer from synthetic data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9194–9204, 2022.

Mousavi, S. M. and Beroza, G. C. Deep-learning seismology. *Science*, 377(6607):eabm4470, 2022. doi: 10.1126/science.abm4470.

Mousavi, S. M. and Beroza, G. C. Machine learning in earthquake seismology. *Annual Review of Earth and Planetary Sciences*, 51 (1):105–129, 2023. doi: 10.1146/annurev-earth-071822-100323.

Mousavi, S. M. and Langston, C. A. Automatic noise-removal/signal-removal based on general cross-validation thresholding in synchrosqueezed domain and its application on earthquake data. *Geophysics*, 82(4):V211–V227, 2017. doi: 10.1190/geo2016-0433.1.

Mousavi, S. M., Sheng, Y., Zhu, W., and Beroza, G. C. STanford EArthquake Dataset (STEAD): A global data set of seismic signals for AI. *IEEE Access*, 7:179464–179476, 2019a. doi: 10.1109/ACCESS.2019.2947848.

Mousavi, S. M., Zhu, W., Sheng, Y., and Beroza, G. C. CRED: A deep residual network of convolutional and recurrent units for earthquake signal detection. *Scientific reports*, 9(1):10267, 2019b. doi: 10.1038/s41598-019-45748-1.

Mousavi, S. M., Ellsworth, W. L., Zhu, W., Chuang, L. Y., and Beroza, G. C. Earthquake transformer—an attentive deep-learning model for simultaneous earthquake detection and phase picking. *Nature communications*, 11(1):3952, 2020. doi: 10.1038/s41467-020-17591-w.

Münchmeyer, J., Woollam, J., Rietbrock, A., Tilmann, F., Lange, D., Bornstein, T., Diehl, T., Giunchi, C., Haslinger, F., Jozinović, D., et al. Which picker fits my data? A quantitative evaluation of deep learning based seismic pickers. *Journal of Geophysical Research: Solid Earth*, 127(1):e2021JB023499, 2022. doi: 10.1029/2021JB023499.

Nakano, M., Sugiyama, D., Hori, T., Kuwatani, T., and Tsuboi, S. Discrimination of seismic signals from earthquakes and tectonic tremor by applying a convolutional neural network to running spectral images. *Seismological Research Letters*, 90(2A): 530–538, 2019. doi: 10.1785/0220180279.

Napoli, A. and White, P. R. Unsupervised domain adaptation for the cross-dataset detection of humpback whale calls. *Detection and Classification of Acoustic Scenes and Events*, 2023.

Negi, S. S., Kumar, A., Ningthoujam, L. S., and Pandey, D. K. An Efficient Approach of Data Adaptive Polarization Filter to Extract Teleseismic Phases from the Ocean-Bottom Seismograms. *Seismological Society of America*, 92(1):528–542, 2021. doi: 10.1785/0220200034.

Nettles, M. and Ekström, G. Glacial earthquakes in Greenland and Antarctica. *Annual Review of Earth and Planetary Sciences*, 38: 467–491, 2010. doi: 10.1146/annurev-earth-040809-152414.

Niksejel, A. and Zhang, M. OBSTransformer: a deep-learning seismic phase picker for OBS data using automated labelling and transfer learning. *Geophysical Journal International*, 237(1): 485–505, 2024. doi: 10.1093/gji/ggae049.

O'Neel, S., Marshall, H. P., McNamara, D. E., and Pfeffer, W. T. Seismic detection and analysis of icequakes at Columbia Glacier, Alaska. *Journal of Geophysical Research: Earth Surface*, 112(F3), 2007. doi: 10.1029/2006JF000595.

Pakhomov, A. and Goldburt, T. Seismic systems for unconventional target detection and identification. In *Sensors, and Command, Control, Communications, and Intelligence (C3I) Technologies for Homeland Security and Homeland Defense V*, volume 6201, pages 466–477. SPIE, 2006. doi: 10.1117/12.668930.

Peláez-Vegas, A., Mesejo, P., and Luengo, J. A survey on semi-supervised semantic segmentation. *arXiv preprint arXiv:2302.09899*, 2023. doi: 10.48550/arXiv.2302.09899.

Peng, L., Li, L., Mousavi, S. M., Zeng, X., and Beroza, G. C. TwoStream-EQT: A microseismic phase picking model combining time and frequency domain inputs. *Computers & Geosciences*, page 105991, 2025. doi: 10.1016/j.cageo.2025.105991.

Pereira, A., Harris, D., Tyack, P., and Matias, L. On the use of the Lloyd's Mirror effect to infer the depth of vocalizing fin whales. *The Journal of the Acoustical Society of America*, 148(5): 3086–3101, 2020. doi: 10.1121/10.0002426.

Piczak, K. J. Environmental sound classification with convolutional neural networks. In *2015 IEEE 25th international workshop on machine learning for signal processing (MLSP)*, pages 1–6. IEEE, 2015. doi: 10.1109/MLSP.2015.7324337.

Plourde, A. P. and Nedimović, M. R. Monitoring fin and blue whales in the lower St. Lawrence Seaway with onshore seismometers. *Remote Sensing in Ecology and Conservation*, 8(4): 551–563, 2022. doi: 10.1002/rse2.261.

Rasmussen, J. H. and Širović, A. Automatic detection and classification of baleen whale social calls using convolutional neural networks. *The Journal of the Acoustical Society of America*, 149 (5):3635–3644, 2021. doi: 10.1121/10.0005047.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

Roch, M. A., Scott Brandes, T., Patel, B., Barkley, Y., Baumann-Pickering, S., and Soldevilla, M. S. Automated extraction of odontocete whistle contours. *The Journal of the Acoustical Society of America*, 130(4):2212–2223, 2011. doi: 10.1121/1.3624821.

Romagosa, M., Baumgartner, M., Cascão, I., Lammers, M. O., Marques, T. A., Santos, R. S., and Silva, M. A. Baleen whale acous-

tic presence and behaviour at a Mid-Atlantic migratory habitat, the Azores Archipelago. *Scientific Reports*, 10(1):4766, 2020. doi: 10.1038/s41598-020-61849-8.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. doi: 10.1007/978-3-319-24574-4$_2$8.

Saad, O. M., Huang, G., Chen, Y., Savvaidis, A., Fomel, S., Pham, N., and Chen, Y. Scalodeep: A highly generalized deep learning framework for real-time earthquake detection. *Journal of Geophysical Research: Solid Earth*, 126(4):e2020JB021473, 2021. doi: 10.1029/2020JB021473.

Samaran, F., Adam, O., and Guinet, C. Detection range modeling of blue whale calls in Southwestern Indian Ocean. *Applied Acoustics*, 71(11):1099–1106, 2010. doi: 10.1016/j.apacoust.2010.05.014.

Saoulis, A. A., Loureiro, A., Tsekhmistrenko, M., and Ferreira, A. M. `reverb`: software, 2025. doi: 10.5281/zenodo.17515563. v1.0.2.

Shakeel, M., Nishida, K., Itoyama, K., and Nakadai, K. 3d convolution recurrent neural networks for multi-label earthquake magnitude classification. *Applied Sciences*, 12(4):2195, 2022. doi: 10.3390/app12042195.

Sharif Razavian, A., Azizpour, H., Sullivan, J., and Carlsson, S. CNN features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 806–813, 2014.

Si, X., Wu, X., Sheng, H., Zhu, J., and Li, Z. SeisCLIP: A seismology foundation model pre-trained by multimodal data for multipurpose seismic feature extraction. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–13, 2024. doi: 10.1109/TGRS.2024.3354456.

Širović, A., Hildebrand, J. A., Wiggins, S. M., McDonald, M. A., Moore, S. E., and Thiele, D. Seasonality of blue and fin whale calls and the influence of sea ice in the Western Antarctic Peninsula. *Deep Sea Research Part II: Topical Studies in Oceanography*, 51(17-19):2327–2344, 2004. doi: https://doi.org/10.1016/j.dsr2.2004.08.005.

Širović, A., Hildebrand, J. A., and Wiggins, S. M. Blue and fin whale call source levels and propagation range in the Southern Ocean. *The Journal of the Acoustical Society of America*, 122(2): 1208–1215, 2007. doi: 10.1121/1.2749452.

Skop, R. and Griffin, O. On a theory for the vortex-excited oscillations of flexible cylindrical structures. *Journal of Sound and Vibration*, 41(3):263–274, August 1975. doi: 10.1016/s0022-460x(75)80173-8.

Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C. A., Cubuk, E. D., Kurakin, A., and Li, C.-L. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

Stafford, K. M., Nieukirk, S. L., Cox, C. G., et al. Geographic and seasonal variation of blue whale calls in the North Pacific. *J. Cetacean Res. Manage.*, 3(1):65–76, 2001. doi: 10.47536/jcrm.v3i1.902.

Stähler, S. C., Sigloch, K., Hosseini, K., Crawford, W. C., Barruol, G., Schmidt-Aursch, M. C., Tsekhmistrenko, M., Scholz, J.-R., Mazzullo, A., and Deen, M. Performance report of the RHUM-RUM ocean bottom seismometer network around La Réunion, western Indian Ocean. *Advances in Geosciences*, 41:43–63, 2016. doi: 10.5194/adgeo-41-43-2016.

Stepnov, A., Chernykh, V., and Konovalov, A. The seismo-performer: a novel machine learning approach for general and efficient seismic phase recognition from local earthquakes in real time. *Sensors*, 21(18):6290, 2021. doi: 10.3390/s21186290.

Storchak, D. A., Di Giacomo, D., Bondár, I., Engdahl, E. R., Harris, J., Lee, W. H., Villaseñor, A., and Bormann, P. Public release of the ISC–GEM global instrumental earthquake catalogue (1900–2009). *Seismological Research Letters*, 84(5):810–815, 2013. doi: 10.1785/0220130034.

Stowell, D. Computational bioacoustics with deep learning: a review and roadmap. *PeerJ*, 10:e13152, 2022. doi: 10.7717/peerj.13152.

Strudel, R., Garcia, R., Laptev, I., and Schmid, C. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7262–7272, 2021.

Stähler, S. C., Schmidt-Aursch, M. C., Hein, G., and Mars, R. A Self-Noise Model for the German DEPAS OBS Pool. *Seismological Research Letters*, 89(5):1838–1845, 2018. doi: 10.1785/0220180056.

Sutton, G. H. and Duennebier, F. K. Optimum design of ocean bottom seismometers. *Marine geophysical researches*, 9:47–65, 1987. doi: 10.1007/BF00338250.

Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., and Liang, J. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312, 2016. doi: 10.1109/TMI.2016.2535302.

Tan, D., Fee, D., Witsil, A., Girona, T., Haney, M., Wech, A., Waythomas, C., and Lopez, T. Detection and characterization of seismic and acoustic signals at Pavlof Volcano, Alaska, using deep learning. *Journal of Geophysical Research: Solid Earth*, 129 (6):e2024JB029194, 2024. doi: 10.1029/2024JB029194.

Tan, T., Godin, O. A., Walters, M. W., and Joseph, J. E. Physics-informed and machine learning-enabled retrieval of ocean current speed from flow noise. *The Journal of the Acoustical Society of America*, 157(2):1084–1096, 2025. doi: 10.1121/10.0035800.

Tarakanov, R. Y., Morozov, E. G., van Haren, H., Makarenko, N. I., and Demidova, T. A. Structure of the Deep Spillway in the Western Part of the Romanche Fracture Zone. *Journal of Geophysical Research: Oceans*, 123(11):8508–8531, 2018. doi: 10.1029/2018jc013961.

Tarvainen, A. and Valpola, H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

Terven, J., Córdova-Esparza, D.-M., and Romero-González, J.-A. A comprehensive review of yolo architectures in computer vision: From yolov1 to yolov8 and yolo-nas. *Machine learning and knowledge extraction*, 5(4):1680–1716, 2023. doi: 10.3390/make5040083.

Tkachenko, M., Malyuk, M., Holmanyuk, A., and Liubimov, N. Label Studio: Data labeling software, 2020-2024. https://github.com/HumanSignal/label-studio. Open source software available from https://github.com/HumanSignal/label-studio.

Trabattoni, A., Barruol, G., Dréo, R., and Boudraa, A. Ship detection and tracking from single ocean-bottom seismic and hydroacoustic stations. *The Journal of the Acoustical Society of America*, 153(1):260–273, 2023. doi: 10.1121/10.0016810.

Trappolini, D., Laurenti, L., Poggiali, G., Tinti, E., Galasso, F., Michelini, A., and Marone, C. Cold diffusion model for seismic denoising. *Journal of Geophysical Research: Machine Learning and Computation*, 1(2):e2024JH000179, 2024. doi: 10.1029/2024JH000179.

Trehu, A. A note on the effect of bottom currents on an ocean bottom seismometer. *Bulletin of the Seismological Society of America*, 75(4):1195–1204, 08 1985a. doi: 10.1785/BSSA0750041195.

Trehu, A. M. Coupling of ocean bottom seismometers to sediment: Results of tests with the U.S. Geological Survey ocean bottom seismometer. *Bulletin of the Seismological Society of America*, 75(1):271–289, February 1985b. doi: 10.1785/bssa0750010271.

Triantafyllou, M. S., Bourguet, R., Dahl, J., and Modarres-Sadeghi, Y. *Vortex-Induced Vibrations*, pages 819–850. Springer International Publishing, 2016. doi: 10.1007/978-3-319-16649-0_36.

Tsekhmistrenko, M., Ferreira, A. M., Miranda, M., Baranbooei, S., Cabieces Diaz, R., Carapuço, M., Corela, C., Duarte, J. L., Ferreira, H., Geissler, W. H., Harris, K., Hicks, S. P., Hosseini, K., Ke, K.-Y., Krüger, F., Lange, D., Loureiro, A., Makus, P., Marignier, A., Neres, M., Ramos, L., Rein, T., Saoulis, A., Schlaphorst, D., Schmidt-Aursch, M. C., and Tilmann, F. Performance of the 2021-2022 UPFLOW large ocean bottom seismometer array in the Azores-Madeira-Canary Islands region, Atlantic Ocean. *Seismica*, 2025.

UIDB/50019/2020, 2020. doi: 10.54499/UIDB/50019/2020.

UIDP/50019/2020, 2020. doi: 10.54499/UIDP/50019/2020.

Valente, R., Correia, A. M., Gil, A., Gonzalez Garcia, L., and Sousa-Pinto, I. Baleen whales in Macaronesia: occurrence patterns revealed through a bibliographic review. *Mammal Review*, 49(2): 129–151, 2019. doi: 10.1111/mam.12148.

Van den Oord, A., Dieleman, S., and Schrauwen, B. Deep content-based music recommendation. *Advances in neural information processing systems*, 26, 2013.

Van Engelen, J. E. and Hoos, H. H. A survey on semi-supervised learning. *Machine learning*, 109(2):373–440, 2020. doi: 10.1007/s10994-019-05855-6.

van Geel, N. C. F., Merchant, N. D., Culloch, R. M., Edwards, E. W. J., Davies, I. M., O'Hara Murray, R. B., and Brookes, K. L. Exclusion of tidal influence on ambient sound measurements. *The Journal of the Acoustical Society of America*, 148(2):701–712, August 2020. doi: 10.1121/10.0001704.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Venkatesh, S., Moffat, D., and Miranda, E. R. You only hear once: a YOLO-like algorithm for audio segmentation and sound event detection. *Applied Sciences*, 12(7):3293, 2022. doi: 10.3390/app12073293.

Walter, F., O'Neel, S., McNamara, D., Pfeffer, W., Bassis, J. N., and Fricker, H. A. Iceberg calving during transition from grounded to floating ice: Columbia Glacier, Alaska. *Geophysical Research Letters*, 37(15), 2010. doi: 10.1029/2010GL043201.

Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., et al. Yolov10: Real-time end-to-end object detection. *Advances in Neural Information Processing Systems*, 37:107984–108011, 2024.

Wang, T., Bian, Y., Zhang, Y., and Hou, X. Using artificial intelligence methods to classify different seismic events. *Seismological Society of America*, 94(1):1–16, 2023. doi: 10.1785/0220220055.

Webb, S. C. Broadband seismology and noise under the ocean. *Reviews of Geophysics*, 36(1):105–142, 02 1998. doi: 10.1029/97rg02287.

Wei, Y., Liang, X., Chen, Y., Shen, X., Cheng, M.-M., Feng, J., Zhao, Y., and Yan, S. Stc: A simple to complex framework for weakly-supervised semantic segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(11):2314–2320, 2016. doi: 10.1109/TPAMI.2016.2636150.

Wilcock, W. S. Tracking fin whales in the northeast Pacific Ocean with a seafloor seismic network. *The Journal of the Acoustical Society of America*, 132(4):2408–2419, 2012. doi: 10.1121/1.4747017.

Wilcock, W. S. and Hilmo, R. S. A method for tracking blue whales (Balaenoptera musculus) with a widely spaced network of ocean bottom seismometers. *Plos one*, 16(12):e0260273, 2021. doi: 10.1371/journal.pone.0260273.

Woollam, J., Münchmeyer, J., Tilmann, F., Rietbrock, A., Lange, D., Bornstein, T., Diehl, T., Giunchi, C., Haslinger, F., Jozinović, D., et al. SeisBench—A toolbox for machine learning in seismology. *Seismological Society of America*, 93(3):1695–1709, 2022. doi: 10.1785/0220210324.

Wu, Y., Yang, T., Liu, D., Dai, Y., and An, C. Current-induced noise in ocean bottom seismic data: Insights from a laboratory water flume experiment. *Earth and Space Science*, 10(6):e2022EA002531, 2023. doi: https://doi.org/10.1029/2022EA002531.

Xi, Z., Wei, S. S., Zhu, W., Beroza, G. C., Jie, Y., and Saloor, N. Deep Learning for Deep Earthquakes: Insights from OBS Observations of the Tonga Subduction Zone. *Geophysical Journal International*, 238(2):ggae200, 2024. doi: 10.1093/gji/ggae200.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., and Luo, P. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

Yang, X., Song, Z., King, I., and Xu, Z. A survey on deep semi-supervised learning. *IEEE transactions on knowledge and data engineering*, 35(9):8934–8954, 2022. doi: 10.1109/TKDE.2022.3220219.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014.

Yu, S., Ma, J., and Wang, W. Deep learning for denoising. *Geophysics*, 84(6):V333–V350, 2019. doi: 10.1190/geo2018-0668.1.

Yu, X., Wang, J., Zhao, Y., and Gao, Y. Mix-ViT: Mixing attentive vision transformer for ultra-fine-grained visual categorization. *Pattern Recognition*, 135:109131, 2023. doi: 10.1016/j.patcog.2022.109131.

Zali, Z., Rein, T., Krüger, F., Ohrnberger, M., and Scherbaum, F. Ocean bottom seismometer (OBS) noise reduction from horizontal and vertical components using harmonic–percussive separation algorithms. *Solid Earth*, 14(2):181–195, 2023. doi: 10.5194/se-14-181-2023.

Zhai, X., Kolesnikov, A., Houlsby, N., and Beyer, L. Scaling vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12104–12113, June 2022.

Zhong, Y. and Tan, Y. J. Deep-Learning-Based Phase Picking for Volcano-Tectonic and Long-Period Earthquakes. *Geophysical Research Letters*, 51(12):e2024GL108438, 2024. doi: 10.1029/2024GL108438.

Zhou, T., Zhang, F., Chang, B., Wang, W., Yuan, Y., Konukoglu, E., and Cremers, D. Image Segmentation in Foundation Model Era: A Survey. *CoRR*, 2024.

Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., and Liang, J. Unet++: A nested u-net architecture for medical image segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support: 4th international workshop, DLMIA 2018, and 8th international workshop, ML-CDS 2018, held in conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, proceedings 4*, pages 3–11. Springer, 2018.

Zhou, Z.-H. A brief introduction to weakly supervised learning. *National Science Review*, 5(1):44–53, 08 2017. doi: 10.1093/n-

sr/nwx106.

Zhu, J., Fang, L., Miao, F., Fan, L., Zhang, J., and Li, Z. Deep learning and transfer learning of earthquake and quarry-blast discrimination: applications to southern california and eastern kentucky. *Geophysical Journal International*, 236(2):979–993, 11 2023. doi: 10.1093/gji/ggad463.

Zhu, W. and Beroza, G. C. PhaseNet: a deep-neural-network-based seismic arrival-time picking method. *Geophysical Journal International*, 216(1):261–273, 10 2018. doi: 10.1093/gji/ggy423.

Zhu, W., Mousavi, S. M., and Beroza, G. C. Seismic signal denoising and decomposition using deep neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 57(11):9476–9488, 2019. doi: 10.1109/TGRS.2019.2926772.

Zou, Z., Chen, K., Shi, Z., Guo, Y., and Ye, J. Object detection in 20 years: A survey. *Proceedings of the IEEE*, 111(3):257–276, 2023. doi: 10.1109/JPROC.2023.3238524.