IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 25 June 2026, accepted 11 July 2026. Date of publication 00 xxxx 0000, date of current version 00 xxxx 0000.

*Digital Object Identifier 10.1109/ACCESS.2026.3714099*

# Marine Mammal Acoustic Species Recognition via Metric Embedding Learning With Audio Spectrogram Transformer

**VINCENZO LIPARI, (Member, IEEE), GIUSEPPE BRANCATELLI, AND EDY FORLIN**
National Institute of Oceanography and Applied Geophysics (OGS), Sgonico, 34010 Trieste, Italy
Corresponding author: Vincenzo Lipari (vlipari@ogs.it)

**ABSTRACT** Automated classification of marine mammal vocalizations from passive acoustic monitoring data is essential for ecological research, biodiversity assessment, and the management of anthropogenic impacts on marine ecosystems. Existing deep learning approaches achieve high accuracy on predefined species sets but operate within a closed-set paradigm: they cannot accommodate previously unseen species without complete retraining, which limits their practical deployment in the field. In this work, we propose a metric embedding framework for marine mammal acoustic species classification. Our approach combines an Audio Spectrogram Transformer (AST) backbone, pre-trained on the large-scale AudioSet dataset, with a projection head trained using a combined ArcFace and online triplet loss. The model learns a 256-dimensional embedding space in which species similarity is captured by cosine distance, enabling closed-set classification, open-set detection of unseen species, and few-shot learning of novel classes from a limited number of reference recordings. We evaluate our method on the Watkins Marine Mammal Sound Database across 25 known species and 6 species held out for open-set and few-shot evaluation. Averaged over three independent training runs, the model attains $92.4 \pm 0.5\%$ closed-set accuracy and a macro F1-score of $0.926 \pm 0.005$, statistically indistinguishable from a fine-tuned softmax classifier on the same backbone (McNemar test, $p = 0.625$). The same embedding additionally detects unseen species with an AUROC of $0.891 \pm 0.026$ under a genuine leave-N-species-out protocol, and discovers new species with a macro F1-score of 0.791 at $K = 10$ under a leakage-free, validation-tuned threshold. These results demonstrate that metric embedding learning provides a flexible and scalable paradigm for marine mammal acoustic monitoring, supporting the integration of new species into deployed systems without retraining.

**INDEX TERMS** Marine mammal acoustics, metric learning, audio embeddings, few-shot classification, open-set classification, audio spectrogram transformer, bioacoustics, species identification.

## I. INTRODUCTION

Marine mammal vocalizations play a fundamental role in communication, navigation, and social interaction within marine ecosystems. Passive acoustic monitoring (PAM) has become a primary tool for studying marine mammal populations, behavior, and the impact of anthropogenic noise, and generates massive volumes of underwater acoustic data. However, the scale of the data precludes extensive manual analysis. Automated classification systems capable of reliably identifying marine mammal species from their vocalizations are therefore essential, both for ecological research and for regulatory compliance in marine operations [1].

Existing approaches to marine mammal call classification have evolved through three stages. Early methods relied on hand-crafted features combined with rule-based or manual classification [2], [3]. Traditional machine learning classifiers, including Support Vector Machines (SVMs) and Gaussian Mixture Models (GMMs), improved automation but required careful feature engineering and

The associate editor coordinating the review of this manuscript and approving it for publication was Manuel Rosa-Zurera.

domain expertise [4], [5]. More recently, deep learning models - Convolutional Neural Networks (CNNs), Long Short-Term Memory Networks (LSTMs), and hybrid architectures - have achieved strong performance by learning features directly from spectrograms [6], [7]. Transfer learning from large-scale audio datasets such as AudioSet [8] has further improved results under limited data conditions [9], [10].

However, most existing deep learning approaches for marine mammal call classification operate under a closed-set paradigm: they are trained to discriminate among a fixed set of predefined species and cannot accommodate previously unseen species without collecting new labeled data and retraining the entire model from scratch. This limitation severely restricts their practical deployment in dynamic field conditions, where new species, rare vocalizations, or changing ecosystems are frequently encountered.

This closed-set assumption is increasingly at odds with how passive acoustic monitoring is actually used. Field deployments routinely encounter vocalizations from species outside the training inventory, and the cost of overlooking them—a range shift, an unexpected migrant, a poorly cataloged population—is precisely what monitoring is meant to detect. We therefore argue that the recognition task should be reframed: rather than asking only *which of these known species is this?*, a deployable system must also answer *is this a species I know at all?* and *can I learn a new one from a handful of examples?* The present work treats these three abilities—closed-set recognition, open-set rejection, and few-shot discovery—as a single problem to be solved by one embedding, rather than as separate models bolted together.

An alternative paradigm, developed in the computer vision community for face recognition, is metric embedding learning [11], [12]. Rather than training a classifier to predict discrete species labels, a metric learning model learns to map inputs into a continuous embedding space where the distance between representations reflects semantic similarity. In this space, samples from the same class are mapped to nearby points, while samples from different classes are pushed apart. Classification then reduces to a nearest-neighbor search: a new recording is compared against a small set of reference embeddings, and the closest match determines the species identity. Crucially, this formulation decouples the model from any fixed species inventory. Adding a previously unseen species requires only a few reference recordings — not model retraining — making the approach naturally suited to open-set and few-shot settings [13].

Metric learning has been successfully applied to the visual identification of individual marine mammals from dorsal fin photographs [14]. In the acoustic domain, contrastive and prototypical learning approaches have been explored for general bioacoustic event detection, particularly within the DCASE few-shot challenges [15], [16], demonstrating the potential of metric-based representations in low-data regimes. To the best of our knowledge, we have not identified prior work combining Audio Spectrogram Transformers with angular-margin objectives and triplet-based metric learning for multi-species marine mammal acoustic classification in open-set and few-shot settings.

This work introduces a metric embedding framework for marine mammal species classification based on an Audio Spectrogram Transformer (AST) backbone [17] pre-trained on AudioSet [8]. The AST features are projected into a compact 256-dimensional $\ell^2$-normalized embedding space, trained with a combination of ArcFace loss (for angular separation between species clusters) and online triplet loss with semi-hard negative mining (for local distance refinement).

At inference time, species identification is performed by comparing query embeddings to species prototypes via cosine distance, enabling few-shot classification of species not seen during training.

Our main contributions are as follows:

1) We propose a single metric embedding that supports closed-set recognition, open-set detection of unseen species, and few-shot discovery of new species from one model, without retraining.
2) We provide a systematic comparison showing that the proposed metric-learning strategy achieves competitive results across all three capabilities, while competing methods are typically tailored to address only one of the three tasks.
3) We illustrate a training strategy combining an AudioSet-pretrained AST backbone with ArcFace and online triplet loss, adapted to small, imbalanced bioacoustic data.
4) We discuss the impact of dataset splitting strategies on reported performance in marine bioacoustics, highlighting the trade-off between strict cross-session evaluation and statistical reliability when working with curated databases of limited recording diversity.

## II. RELATED WORK

Automatic marine mammal call recognition and classification lies at the intersection of bioacoustics, signal processing, and machine learning. In this section, we review the main lines of research relevant to our work: acoustic classification methods for marine mammals, metric learning and embedding approaches, few-shot learning for bioacoustics, and audio transformer architectures.

### A. MARINE MAMMALS ACOUSTIC CLASSIFICATION

Early approaches to marine mammal call classification relied on hand-crafted acoustic features followed by manual or rule-based classification. Reference [2] applied the ISHMAEL and PAMGUARD [18] systems for passive acoustic detection and manual classification of six marine mammal species. Reference [3] proposed generalized perceptual linear prediction (GPLP) and Greenwood function cepstral coefficients (GFCC) as species-independent feature representations. While these methods demonstrated the feasibility of acoustic species identification, they required substantial manual intervention and expert knowledge,

and their accuracy degraded when different species produced acoustically similar vocalizations.

The introduction of machine learning (ML) classifiers enabled greater automation. Reference [4] demonstrated that combining low-frequency cepstral coefficients with discrete wavelet transforms could effectively characterize the upcalls of North Atlantic right whales when paired with Support Vector Machines (SVM). Subsequent studies validated that the fusion of multiple feature domains—specifically Mel-frequency cepstral coefficients (MFCC), linear frequency descriptors, and time-domain features—yielded consistent performance gains [5]. Despite these advancements, these paradigms remain inherently limited by the necessity of manual feature engineering, which scales poorly to global-scale datasets and struggles to generalize across heterogeneous recording conditions.

Deep learning approaches have substantially advanced the state of the art by learning features directly from spectral representations. Reference [19] proposed a CNN–GRU architecture to jointly extract acoustic features and perform temporal modeling. Reference [6] introduced a multi-channel parallel network combining Mel spectrograms and complementary features. Reference [20] applied residual learning networks to classify calls. More recently, [10] proposed a hybrid architecture combining time-attention LSTMs with multi-scale dilated causal convolutions.

Transfer learning from large-scale audio datasets has emerged as a key strategy for addressing data scarcity in marine bioacoustics. Reference [9] employed transfer learning with deep convolutional networks to detect multiple marine sound sources, achieving high-accuracy identification of odontocete whistles and ship noise. Reference [21] proposed a framework based on knowledge distillation for efficient marine mammal sound detection. In a different direction, [22] leveraged a speech self-supervised model pretrained on large amounts of human voice data and fine-tuned it for marine mammal classification, demonstrating that representations learned from human speech transfer effectively to marine vocalizations.

### B. METRIC LEARNING AND EMBEDDING APPROACHES

Metric learning addresses the limitations of closed-set classifiers by learning an embedding function that maps inputs into a space where semantic similarity is captured by distance. Rather than predicting class labels directly, the model learns to position similar inputs close together and dissimilar inputs far apart. At inference time, classification is performed by comparing a query embedding to reference examples using a distance metric, naturally supporting open-set recognition and few-shot learning.

The triplet loss, introduced by [11] in the FaceNet architecture for face recognition, trains the model using triplets of samples: an anchor, a positive example (same class), and a negative example (different class). The loss encourages the anchor-positive distance to be smaller than the anchor-negative distance by a margin.

Angular margin losses [12] extend the metric learning paradigm by operating on the angular space of $\ell^2$-normalized embeddings. These angular margin methods have been shown to produce more discriminative embedding spaces than triplet loss alone.

In marine biology, metric learning has been applied to visual identification of individual animals. Reference [14] used a ResNet backbone trained with triplet loss to learn a Euclidean embedding of common dolphin dorsal fin images, achieving strong identification performance even for individuals not seen during training. These results demonstrate the viability of metric learning for marine mammal identification, but the approach has been limited to visual data.

### C. AUDIO SPECTROGRAM TRANSFORMERS

Vision Transformers (ViTs), originally designed for image classification [23], have been successfully adapted to audio analysis by treating spectrograms as images. The Audio Spectrogram Transformer (AST) [17] divides a log-mel spectrogram into a sequence of $16\times16$ patches, applies learned positional embeddings, and processes the patch sequence through a standard transformer encoder. Pre-trained on AudioSet [8], a large-scale dataset of audio clips, AST achieved state-of-the-art results on multiple audio classification benchmarks, demonstrating that the self-attention mechanism can effectively capture both local spectral patterns and long-range temporal dependencies in audio data.

The transfer of pre-trained audio representations to bioacoustic tasks has shown promising results. Foundation models pre-trained on AudioSet or speech data have been applied to bird song classification, whale call detection, and general environmental sound recognition, often outperforming task-specific models trained from scratch. Recent comparative studies [24] have systematically evaluated bioacoustic foundation models, confirming that pre-trained transformers provide robust feature extractors across diverse taxa and recording conditions.

However, existing applications of audio transformers in marine bioacoustics have used these models as feature extractors for closed-set softmax classifiers.

Beyond task-specific models, several pre-trained audio foundation models have recently emerged. General-audio models such as CLAP [25] learn joint audio–language embeddings from large web corpora, while self-supervised bioacoustic encoders such as AVES [26] are trained directly on animal-sound collections. Domain-specific models—BirdNET [27], the Perch birdsong embeddings [28], and acoustic-token models such as BEATs [29]—have shown strong transfer to downstream bioacoustic tasks, and are typically used as frozen feature extractors. Our work differs in that we fine-tune the backbone with task-specific metric supervision and ask whether a single embedding can support closed-set recognition, open-set detection,

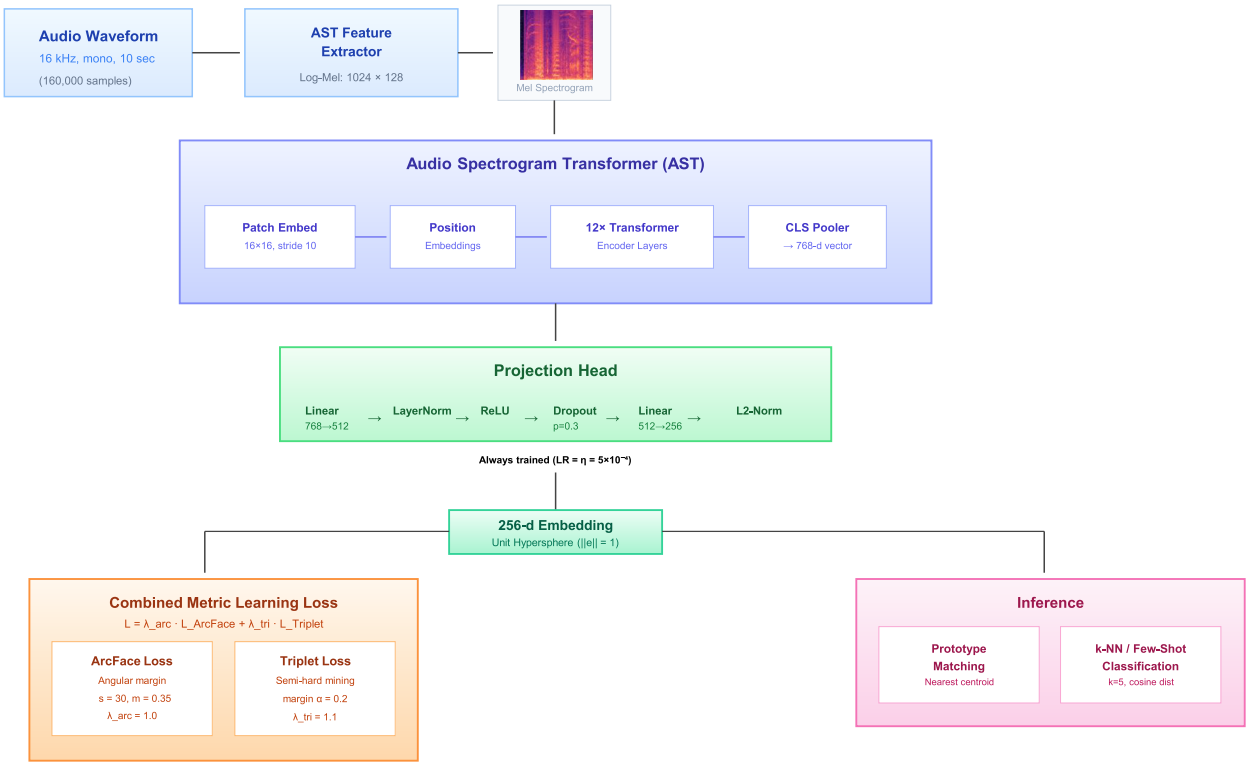


**FIGURE 1.** Overview of the proposed metric embedding architecture for marine mammal species classification.

and few-shot discovery together; in Section V-B we benchmark representative frozen encoders (CLAP, AVES, and our AST backbone frozen) against this requirement.

## III. METHODOLOGY

This section describes the proposed metric embedding framework for marine mammal species classification. We first formulate the problem in terms of embedding learning, then detail the model architecture, loss functions, training strategy, and inference procedures.

### A. PROBLEM FORMULATION

Conventional approaches to marine mammal call classification learn a discriminative mapping $f : X \to \{1, \ldots, C\}$ that assigns each audio clip to one of $C$ predefined species categories, typically by training a neural network with a softmax output layer and cross-entropy loss. This formulation is inherently closed-set: the model can only predict among the $C$ classes observed during training.

We adopt an alternative formulation based on metric embedding learning. Instead of predicting class labels, we learn an embedding function $g : X \to \mathbb{R}^d$ that maps each audio clip into a $d$-dimensional vector space, where the geometric distance between embeddings encodes species similarity. Specifically, we require that embeddings of clips from the same species cluster together, while embeddings of clips from different species are well-separated. At inference time, classification reduces to a nearest-neighbor search in the embedding space: a query embedding is compared against species-level reference embeddings (prototypes), and the species of the closest prototype determines the prediction.

This formulation has three key advantages over closed-set classification. First, it naturally supports open-set recognition, enabling the model to distinguish between known categories and unseen species or noise by identifying them as out-of-distribution samples. Second, it facilitates few-shot learning, as new species prototypes can be accurately defined using only a small number of reference recordings. Third, the embedding space provides an interpretable similarity structure where inter-species distances reflect acoustic relatedness; this can be effectively exploited for both exploratory analysis and anomaly detection.

### B. SYSTEM ARCHITECTURE

The proposed system consists of three components: (1) a pre-trained audio backbone that extracts high-level acoustic representations from spectrograms, (2) a projection head that maps these representations into a compact, normalized embedding space, and (3) a combined metric learning loss function that shapes the geometry of the embedding space during training. We describe each component below.

#### 1) AUDIO BACKBONE: AUDIO SPECTROGRAM TRANSFORMER

We adopt the Audio Spectrogram Transformer (AST) [17] as the acoustic feature extractor. AST adapts the Vision

Transformer architecture to log-mel spectrograms by dividing the spectrogram into patches that are processed by a transformer encoder.

The model is initialized with weights pre-trained on AudioSet [8] and produces a 768-dimensional feature representation extracted from the classification token (CLS).

This pre-training provides the backbone with rich, general-purpose acoustic representations that capture both local spectral patterns (e.g., harmonics, clicks) and long-range temporal dependencies (e.g., call sequences, frequency sweeps) - particularly relevant for marine mammal vocalizations, which vary widely in duration, bandwidth, and temporal structure.

The backbone outputs a 768-dimensional feature vector extracted from the CLS token via a pooling layer.

#### 2) PROJECTION HEAD

The backbone features are passed through a projection head that maps them into a compact embedding space. The projection head is a two-layer multilayer perceptron (MLP) with the following architecture:

$$\boxed{\mathbb{R}^{768}}$$
$$\downarrow$$
$$\text{Linear } 768 \to 512$$
$$\downarrow$$
$$\text{LayerNorm}$$
$$\downarrow$$
$$\text{ReLU}$$
$$\downarrow$$
$$\text{Dropout } (0.3)$$
$$\downarrow$$
$$\text{Linear } 512 \to 256$$
$$\downarrow$$
$$\boxed{\mathbb{R}^{256} \ (\text{L}^2 \text{ normalized})}$$

The final L2 normalization constrains all embeddings to lie on a unit hypersphere in $\mathbb{R}^{256}$, ensuring that cosine similarity and Euclidean distance are equivalent up to a monotonic transformation. This normalization is essential for the angular margin loss described below.

#### 3) COMBINED METRIC LEARNING LOSS

We train the embedding model using a weighted combination of two complementary loss functions: ArcFace loss for angular cluster separation and online triplet loss for local distance refinement.

**ArcFace Loss**: Originally developed for face recognition [12], ArcFace utilizes a learnable weight matrix $W \in \mathbb{R}^{C\times d}$, where each row $w_i$ is a class-specific direction in the embedding space. For an $\ell^2$-normalized embedding $e$ with label $y$, the cosine similarity to each class center is computed as $\cos(\theta_y) = w_y^T e$. The loss function is defined as

$$L_{arc} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{s\cdot\cos(\theta_{y_i}+m)}}{e^{s\cdot\cos(\theta_{y_i}+m)} + \sum_{j=1, j\neq y_i}^{C} e^{s\cdot\cos(\theta_j)}} \quad (1)$$

where:

- the angular Margin $m$ is an additive penalty applied to the angle $\theta$ of the target class,
- the feature scale $s$ is a hyperparameter that scales the cosine similarities,

**Online Triplet Loss with Semi-Hard Mining**: The triplet loss [11] acts on triplets of samples: an anchor $a$, a positive example $p$ from the same class, and a negative example $n$ from a different class. It encourages the anchor to be closer to the positive than to the negative by a margin $\alpha$:

$$L_{tri} = \max(0, d(a,p) - d(a,n) + \alpha) \quad (2)$$

where $d(x, y)$ denotes the Euclidean distance between $x$ and $y$. The effectiveness of triplet loss depends on the selection of informative triplets. We employ semi-hard negative mining: i.e., for each anchor-positive pair, we select the negative that is farther from the anchor than the positive but still within the margin, i.e., $d(a,p) < d(a,n) < d(a,p) + \alpha$. This strategy avoids the instability associated with hard negatives (which can cause training collapse) while providing stronger gradients than random sampling.

**Combined Loss** The total training objective is a weighted combination of the two losses:

$$L = \lambda_{arc} L_{arc} + \lambda_{tri} L_{tri} \quad (3)$$

ArcFace enforces global angular separation between species clusters, while triplet loss refines local neighborhood structure by enforcing relative distance constraints among nearby samples.

The overall architecture for embedding learning and classification is depicted in Figure 1.

### C. TRAINING STRATEGY

Training is performed in two phases, with specialized batch construction and differential learning rates.

#### 1) BALANCED BATCH SAMPLING

Effective metric learning is highly sensitive to batch composition; the triplet loss requires each training batch to contain multiple instances per class to form valid anchor-positive-negative triplets. Naive random sampling frequently generates batches with no positive pairs thus making the triplet loss uninformative.

We employ a balanced class-aware sampling strategy [30]. Each mini-batch is constructed by randomly selecting $K$ species and sampling exactly $M$ instances per species; this ensures $K \cdot M(M-1)$ directed anchor-positive pairs per batch, alongside a vast pool of potential negative triplets. This structured sampling is essential for convergence: without it, the triplet loss gradient remains negligible, as the model could

not find sufficient informative triplets to drive the learning process.

#### 2) PROGRESSIVE BACKBONE UNFREEZING

We adopt a two-phase training schedule to balance the preservation of pre-trained representations with adaptation to the marine mammal domain:

- Phase 1 (i.e.,Epochs 1–10): Frozen backbone. The AST backbone parameters are frozen, and only the projection head and ArcFace weight matrix are trained. This allows the projection head to learn a meaningful initial mapping from the pre-trained feature space to the embedding space, without risk of corrupting the backbone representations. During this phase, the learning rate is set to $\eta = 5 \times 10^{-4}$.
- Phase 2 (i.e.,Epochs 11–100 (max)): Full fine-tuning. The backbone is unfrozen, and training continues with differential learning rates: the projection head reduces the base rate to $\eta = 2 \times 10^{-4}$, while the backbone parameters use a reduced rate of $\eta = 10^{-6}$. This reduction prevents catastrophic forgetting of the general audio representations learned from AudioSet, while allowing gradual adaptation to marine mammal-specific acoustic patterns. Training was performed for up to 100 epochs using early stopping based on validation k-NN accuracy with a patience of 10 epochs.

### D. INFERENCE: CLASSIFICATION VIA EMBEDDING COMPARISON

At inference time, the trained model is used to compute embeddings for both reference recordings and new query recordings. Classification is then performed entirely in the embedding space. We consider three classification strategies of increasing complexity:

#### 1) PROTOTYPE MATCHING

For each species, a prototype embedding is computed as the centroid of all reference embeddings for that species. A query recording is classified to the species whose prototype is nearest (in cosine distance). This method is the simplest and most suitable for deployment, as it requires storing only one prototype vector per species.

#### 2) K-NEAREST NEIGHBORS (K-NN)

The query embedding is compared against all individual reference embeddings, and the species is determined by majority vote among the k nearest neighbors (using cosine distance). This method is more robust to outlier reference recordings but requires storing all individual embeddings. We use $k = 5$ in our experiments.

#### 3) THRESHOLD-BASED VERIFICATION

Rather than classifying among all known species, this mode determines whether a query recording belongs to a specific target species by checking if the cosine distance to the species prototype falls below a threshold $\delta$. This mode is suitable for species-specific monitoring tasks.

Crucially, all three strategies operate on fixed-dimensional embeddings and do not require access to the model parameters or class weights. Adding a previously unseen species requires only computing the embeddings for a small number of reference recordings and either computing a new prototype or adding the embeddings to the reference database. No model retraining or architectural modification is needed, which is the main practical advantage of the metric learning formulation.

#### 4) JOINT KNOWN-CLASS CLASSIFICATION AND NOVEL-CLASS DISCOVERY

For deployment scenarios where both known and novel species may appear, we adopt a dual-threshold strategy. Given a query embedding $\mathbf{e}$ and a set of known-species prototypes $\{\mathbf{p}_k\}_{k=1}^{C}$, let $s_{\max} = \max_k \cos(\mathbf{e}, \mathbf{p}_k)$. If $s_{\max} \geq \tau_{\text{known}}$, the query is assigned to the corresponding known species. If a novel-class reference set is available, we additionally compare against novel-species prototypes built from $K$ reference samples. A novel-class assignment is accepted only if its similarity exceeds $\tau_{\text{discovery}} = \beta \cdot \tau_{\text{known}}$, with $\beta \in (0, 1)$ controlling the conservativeness of novel detection. Queries falling below both thresholds are abstained from.

Both thresholds are derived from the open-set operating point (Section V-C). The known-class threshold $\tau_{\text{known}} = \tau^*$ is the optimal cosine similarity threshold from the Receiver Operating Characteristic (ROC) analysis, tuned only on a held-out validation split of unseen species and applied unchanged to the disjoint test fold ($\tau^* \approx 0.91$, mean across repeats). The discovery threshold $\tau_{\text{discovery}} = \beta \cdot \tau^*$ with $\beta = 0.8$ creates a conservative margin below $\tau^*$ for accepting novel-class assignments. All thresholds operate on cosine similarity (range $[-1, 1]$), where higher values indicate greater proximity to a class prototype. This formulation creates an explicit uncertainty region $[\tau_{\text{discovery}}, \tau_{\text{known}}]$ where the system declines to commit, a desirable property in passive acoustic monitoring (PAM) applications where overconfident misclassifications can compromise downstream ecological analyses.

## IV. EXPERIMENTAL SETUP

### A. DATASET

Our experiments use the Watkins Marine Mammal Sound Database [31], a curated collection of cetacean and pinniped vocalizations widely used as a benchmark in marine bioacoustics. We accessed the database through the HuggingFace Datasets repository [32], which provides 1,357 high-quality clips spanning 32 species. After removing species with fewer than 8 clips, the effective dataset covers 25 species for closed-set evaluation. All recordings are resampled to 16 kHz, converted to mono, and padded or truncated to a fixed duration of 10 seconds (160,000 samples).

We reserve six species entirely as a holdout set for open-set and few-shot evaluation, chosen to span both taxonomic and acoustic diversity: four phylogenetically distant pinnipeds (*Walrus*, *Ross Seal*, *Bearded Seal*, *Leopard Seal*), one rare odontocete (*Narwhal*), and one common delphinid acoustically similar to several training species (*Bottlenose Dolphin*). The holdout deliberately spans a wide range of sample sizes (8–40 clips), allowing us to probe few-shot behavior both for well-represented and for data-scarce novel species.

The holdout species were selected a priori based on taxonomic and acoustic diversity, before any model evaluation, to span the spectrum from phylogenetically distant pinnipeds to acoustically overlapping delphinids. To verify that open-set performance does not depend on this particular choice, Section V-C additionally reports a leave-$N$-species-out experiment over four different six-species holdout sets.

The closed-set partition is divided using a stratified clip-level split with ratios 70%/15%/15%, yielding 823 training, 168 validation, and 197 test clips. The holdout partition (167 clips, 6 species) is used exclusively for open-set and few-shot evaluation.

*Limitations of dataset splitting in marine bioacoustics.* The Watkins database, while curated for quality, contains a limited number of distinct recordings per species (typically 2–7 sessions). Strict session-aware splits prevent any temporal correlation between train and test data but produce highly variable test sets due to small sample sizes per session. Stratified clip-level splits yield more statistically reliable estimates but may introduce weak temporal correlation when multiple clips originate from the same session. We adopt clip-level splitting as the main evaluation protocol, consistent with the majority of bioacoustic classification literature [33], [34]. We note that the open-set and few-shot evaluations, which use entire holdout species never seen during training, are by construction immune to any clip-level leakage and therefore provide the most realistic assessment of the model's generalization.

### B. DATA PREPROCESSING

Audio clips are preprocessed with the same pipeline used during AudioSet pre-training:computing 128-bin log-mel spectrograms, applying per-instance mean and variance normalization, and padding or truncating the time axis to a fixed length of 1024 frames.

### C. EVALUATION PROTOCOL

#### 1) CLOSED-SET CLASSIFICATION

We report accuracy, precision, recall, and macro/weighted F1-score on the test set via prototype matching: for each species, a centroid is computed from all training embeddings, and a query is assigned to the nearest prototype (cosine distance).

#### 2) EMBEDDING QUALITY

We report the average intra-class distance, inter-class distance, and separation ratio (inter/intra). Higher separation ratios indicate more discriminative embeddings.

#### 3) OPEN-SET DETECTION

Open-set capability is evaluated as a binary task: decide whether a query belongs to one of the known species or to an unknown class, using the maximum cosine similarity to known prototypes as the score. Because the rejection threshold $\tau$ must not be tuned on the evaluation data, we split the non-training species into a validation group and a disjoint test group: $\tau$ is selected only on the validation group (maximizing balanced accuracy of the known/unknown decision) and applied unchanged to the test group, repeated over ten random partitions (reported as mean $\pm$ std). As a stricter test of sensitivity to the held-out species, we additionally perform a leave-$N$-species-out experiment: for four different six-species holdout sets (including the original), the model is re-trained from scratch with those species entirely absent and open-set AUROC is measured on them.

#### 4) FEW-SHOT MULTI-SPECIES DISCOVERY

We evaluate joint known/novel classification under the dual-threshold protocol (Section IV-C), building novel prototypes from $K \in \{1, 5, 10\}$ reference samples and reporting macro F1 across holdout species over 500 episodes, with the validation-tuned threshold $\tau$ gating the known/unknown decision.

#### 5) STATISTICAL RELIABILITY

Closed-set metrics are averaged over three runs ($N = 3$), reported as mean $\pm$ std with 95% confidence intervals via the $t$-distribution. Per-species F1 confidence intervals are obtained by bootstrap resampling [35], and confidence intervals on the open-set AUROC by the method of DeLong [36]. Per-species tables report a representative run (best seed).

### D. IMPLEMENTATION DETAILS

All experiments use the AST backbone pre-trained on AudioSet. The projection head is a two-layer MLP ($768 \to 512 \to 256$) with LayerNorm, Rectified Linear Unit (ReLU), dropout 0.3, and $\ell_2$-normalization. The combined loss is $L = L_{\text{ArcFace}} + 1.1 \cdot L_{\text{triplet}}$ with $s = 30$, $m = 0.35$, $\alpha = 0.20$.

Optimization uses AdamW (weight decay $5 \times 10^{-4}$), learning rate $2 \times 10^{-4}$ on the head and $10^{-6}$ on the backbone.

Training runs for max 100 epochs with 3 warmup epochs and cosine decay; the backbone is frozen for the first 10 epochs. Batches of size 16 use class-balanced sampling (4 per class) [30] with online semi-hard negative mining. Gradient clipping (max norm 0.5) is applied throughout. Training is performed on a single NVIDIA RTX A6000 GPU.

**TABLE 1. Training Hyperparameters and model configuration.**

| Parameter | Value |
|---|---|
| *Model Architecture* | |
| Backbone | AST (`finetuned-audioset`) |
| Embedding Dimension | 256 |
| Projection Hidden Dim | 512 |
| Dropout Rate | 0.3 |
| *Optimization & Loss* | |
| Loss Function | ArcFace + 1.1· Triplet (semi-hard) |
| ArcFace Scale / Margin | 30.0 / 0.35 |
| Triplet Margin | 0.20 |
| Optimizer | AdamW (weight decay WD = $5 \times 10^{-4}$) |
| LR (head / backbone) | $2 \cdot 10^{-4}$ / $10^{-6}$ |
| Warmup / Total Epochs | 3 / 100 |
| Backbone Frozen Epochs | 10 |
| Batch Size / Samples per Class | 16 / 4 |
| Gradient Clipping | 0.5 |

**TABLE 2. Classification and open-set detection results, mean ± std over 3 independent training runs (25 species). Open-set AUROC is reported under the leave-*N*-species-out protocol (primary) and the unknown-split protocol (supplement).**

| Metric | Value |
|---|---|
| *Closed-set classification (25 species)* | |
| Test Accuracy | $92.4 \pm 0.5\%$ |
| Macro F1-score | $0.926 \pm 0.005$ |
| Weighted F1-score | $0.923 \pm 0.005$ |
| Mean max cosine similarity | $0.939 \pm 0.009$ |
| *Open-set detection (6 holdout species)* | |
| AUROC (leave-$N$-out) | $0.891 \pm 0.026$ |
| AUROC (unknown-split) | $0.912 \pm 0.034$ |
| Optimal threshold $\tau^*$ (val-tuned) | $0.909 \pm 0.048$ |
| TPR at $\tau^*$ (operating point) | 0.785 |
| FPR at $\tau^*$ (operating point) | 0.173 |
| Separation ratio (inter/intra) | $5.83 \pm 0.67\times$ |

**TABLE 3. Comparison with metric-learning baselines on the same AST backbone, under the identical protocol. Open-set AUROC under the unknown-split protocol shared by all methods. Few-shot at $K$=10 with known/unknown gating. Best per column in bold.**

| Method | Closed-set acc. | Open-set AUROC | Few-shot F1 |
|---|---|---|---|
| Ours | **0.924** | **0.912** | 0.791 |
| SupCon | 0.814 | 0.857 | **0.825** |
| ProtoNet | 0.745 | 0.866 | 0.799 |

**TABLE 4. Comparison with frozen foundation encoders under the identical protocol. Open-set AUROC under the unknown-split protocol; the few-shot column uses a threshold-free $K$-shot $N$-way macro-F1 (no gating). Best per column in bold.**

| Encoder | Closed-set acc. | Open-set AUROC | Few-shot F1 (t.f.) |
|---|---|---|---|
| Ours (fine-tuned) | **0.924** | **0.912** | 0.844 |
| AVES (frozen) | 0.701 | 0.814 | **0.860** |
| AST (frozen) | 0.609 | 0.820 | 0.804 |
| CLAP (frozen) | 0.665 | 0.847 | 0.773 |

## V. RESULTS

### A. CLOSED-SET CLASSIFICATION

Table 2 summarizes the main results averaged over three independent training runs. The proposed metric embedding framework achieves 92.4 ± 0.5% test accuracy and 0.926 ± 0.005 macro F1 over three seeds. The narrow standard deviation across runs (±0.5%) indicates reliable convergence.

Per-species F1 with 95% bootstrap confidence intervals is reported in Appendix VII. Among the 25 evaluated species, 21 achieve F1 $\geq$ 0.85 and 13 reach F1 $\geq$ 0.94. The mean cosine similarity between test embeddings and their assigned prototype is 0.939 ± 0.009 across runs, indicating that confident predictions dominate. Species with fewer than eight test clips have wide confidence intervals, so their individual point estimates—including perfect scores—should be interpreted with caution.

On a per-species basis the metric model and a softmax classifier sharing the same backbone differ on individual species, but a McNemar exact test on the paired test predictions finds no significant difference between the metric model (92.4%) and the softmax baseline (91.9%): $b = 3$, $c = 1$, $p = 0.625$. We therefore do not claim a closed-set advantage. Frozen foundation encoders, by contrast, trail far behind on closed-set accuracy (Table 4), confirming that task-specific fine-tuning—not the pretrained representation alone—drives recognition performance.

Two species warrant discussion. The *Killer Whale* (recall 0.60, $n = 5$) has a small test set combined with high call variability across age classes. The *Sperm Whale* (recall 0.67, $n = 9$) is partially confused with *Walrus* in the open-set analysis, reflecting genuine acoustic similarity between their click trains.

More broadly, the two dominant sources of difficulty are acoustic overlap between phylogenetically or behaviourally similar species (e.g., the click-train similarity linking Sperm Whale and Walrus) and small per-species sample counts, which widen the bootstrap confidence intervals reported and make individual F1 estimates less reliable.

### B. COMPARISON WITH BASELINES AND FOUNDATION MODELS

We compare our model against two representative metric-learning baselines on the same backbone (ProtoNet [37] and supervised contrastive learning, SupCon [38]) and against three frozen pre-trained encoders evaluated identically: CLAP [25] (general-audio), our AudioSet-pretrained AST used frozen, and AVES [26] (a self-supervised bioacoustic encoder, via its official torchaudio wrapper). Open-set AUROC is reported under the unknown-split protocol shared by all methods; the stricter leave-$N$-out value for our model (0.891) is in Table 2. Against the metric-learning baselines (Table 3), our model leads on closed-set accuracy and open-set AUROC and is comparable on few-shot discovery—the only method strong across all three axes. Against the frozen encoders (Table 4), our fine-tuned model dominates on closed-set accuracy and open-set AUROC; on threshold-free few-shot transfer to novel species it is on par with AVES and above the frozen AST and CLAP. Read across all three axes, no single alternative is strong everywhere: the metric

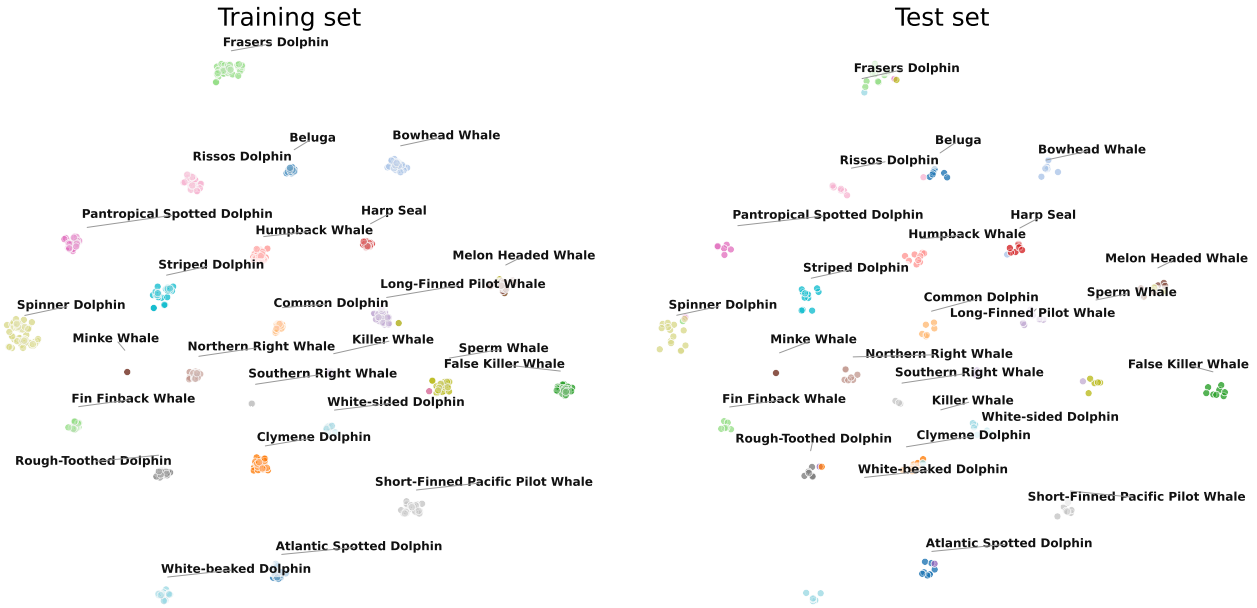


**FIGURE 2.** t-SNE of the learned embedding space for 25 known species. (a) Training set ($n = 823$). (b) Test set ($n = 197$). Projections are computed jointly for spatial comparability. Cluster consistency demonstrates generalization across splits.

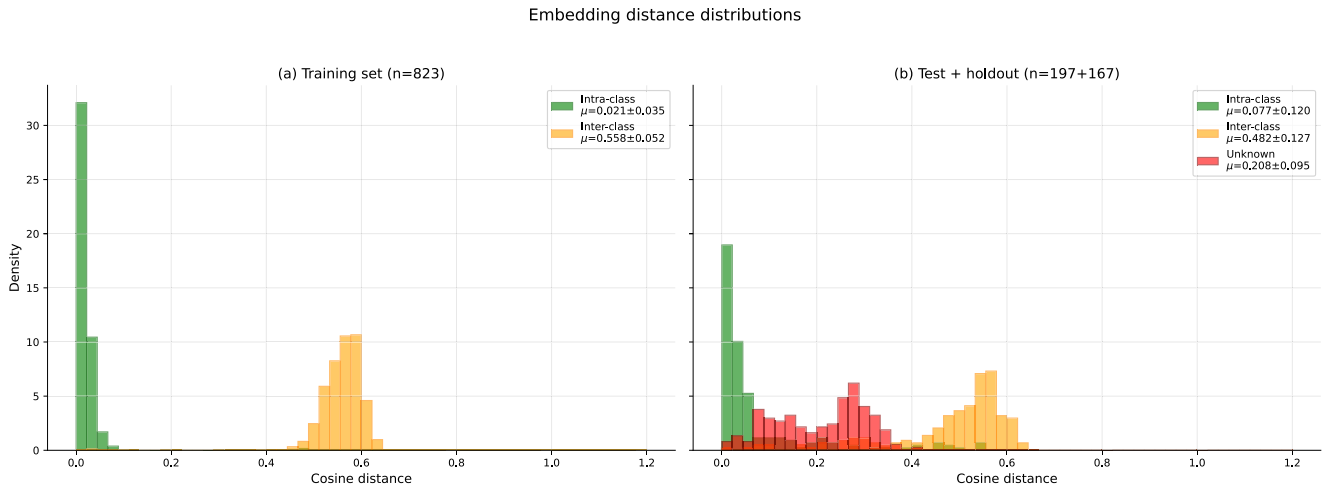


**FIGURE 3.** Cosine distance distributions. (a) Training set: intra- andinter-class distances. (b) Test set and holdout: comparable intra-classdistances confirm no overfitting; unknown distances enablethreshold-based detection.

baselines match us on few-shot but not on closed-set or rejection, and the frozen encoders either collapse on closed-set (CLAP, AST) or recognize known species poorly (AVES, 0.70 vs 0.92). A single fine-tuned embedding covering all three regimes is, to our knowledge, not provided by any off-the-shelf model; large-scale bioacoustic pre-training such as AVES is therefore complementary to our approach.

### C. OPEN-SET DETECTION

Open-set evaluation tests whether the model can distinguish known queries from unknown (holdout) queries, using the maximum cosine similarity to known prototypes as the detection score.

Under the validation-tuned protocol the framework achieves an AUROC of $0.912 \pm 0.034$ over ten random validation/test partitions of the held-out species. At the validation-tuned threshold $\tau^* \approx 0.91$, the model accepts about 78.5% of known queries (TPR) while accepting about 17.3% of unknown queries (FPR). Under the stricter leave-$N$-species-out protocol, in which the model is re-trained from scratch on four different six-species holdout sets, the AUROC remains in a narrow band, $0.891 \pm 0.026$ (range 0.856–0.920, Figure 5); the original holdout sits at the

top of this band (0.920), confirming that the result does not depend on the particular six species chosen. The embedding space achieves a separation ratio of $5.83 \pm 0.67\times$.

Confusions between holdout and known species follow biologically interpretable patterns (Table 8). The *Walrus* is most similar to *Sperm Whale* (similarity 0.60), reflecting shared impulsive click vocalizations. *Bearded Seal* is closest to *Bowhead Whale* (0.55), both producing low-frequency broadband moans. *Bottlenose Dolphin*—selected as a deliberately hard case— shows intermediate similarity to multiple delphinids (max 0.45), confirming that the embedding captures fine-grained acoustic relationships.

### D. FEW-SHOT MULTI-SPECIES DISCOVERY

In realistic PAM deployments, both known and novel species appear simultaneously. We evaluate joint classification using the dual-threshold protocol with the validation-tuned $\tau_{\text{known}} = \tau^* \approx 0.91$ and $\tau_{\text{discovery}} = 0.8 \cdot \tau^*$.

Macro F1 rises monotonically with the number of reference clips, from $0.545 \pm 0.072$ at $K{=}1$ to $0.753 \pm 0.066$ at $K{=}5$ and $0.791 \pm 0.061$ at $K{=}10$ (system metric with the validation-tuned gate active). For a fair encoder-to-encoder comparison we additionally report a threshold-free $K$-shot $N$-way macro-F1 (no gating), on which our embedding scores 0.844, on par with the frozen bioacoustic encoder AVES (0.860, within episode-level variability) and above frozen AST (0.804) and CLAP (0.773); AVES, however, cannot recognize known species without fine-tuning (0.70 vs 0.92), so the two approaches are complementary rather than competing.

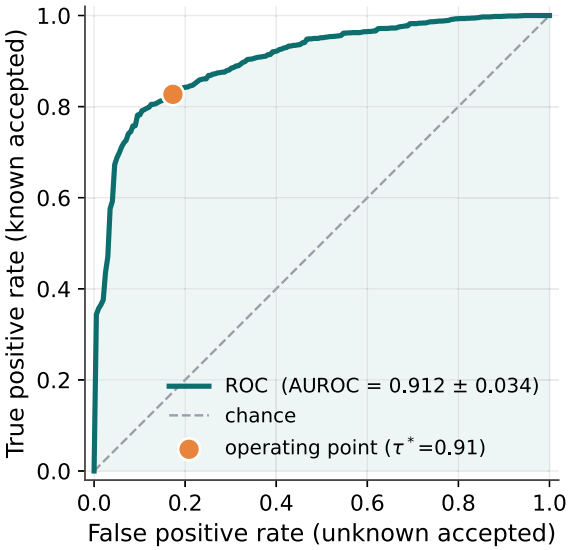


**FIGURE 4.** Open-set ROC under the unknown-split protocol, averaged over ten validation/test partitions (AUROC = 0.912 ± 0.034). The marker is the operating point at the validation-tuned threshold $\tau^* \approx 0.91$; exact TPR/FPR are in table 2. The leave-*N*-species-out AUROC is 0.891 (Sec. V-C).

Figure 6 reports per-species F1 across 500 episodes for $K \in \{1, 5, 10\}$, and Table 5 the macro progression, demonstrating effective generalization to unseen categories

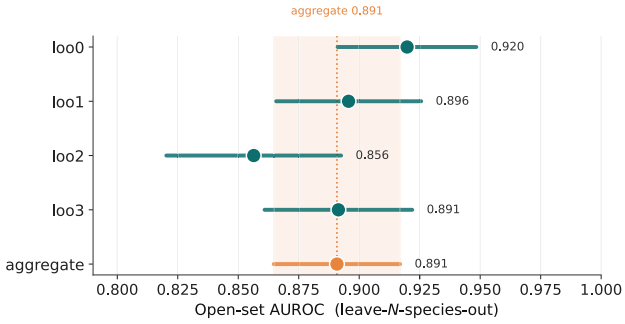


**FIGURE 5.** Open-set AUROC under leave-*N*-species-out: the model is re-trained from scratch on four different six-species holdout sets (including the original), with 95% confidence intervals. Aggregate $0.891 \pm 0.026$.

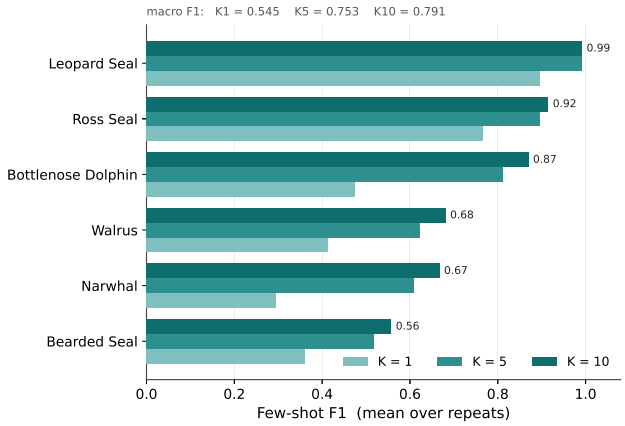


**FIGURE 6.** Few-shot multi-species discovery: per-species F1 at $K \in \{1, 5, 10\}$ (mean over repeats) for the six holdout species. Macro F1 rises from 0.545 ($K{=}1$) to 0.753 ($K{=}5$) to 0.791 ($K{=}10$).

**TABLE 5.** Few-shot multi-species discovery: macro F1 across the six holdout species (500 episodes, mean ± std over repeats), at increasing numbers of reference clips $K$. Per-species F1 is shown in figure 6.

| Reference clips | Macro F1 |
|---|---|
| $K = 1$ | $0.545 \pm 0.072$ |
| $K = 5$ | $0.753 \pm 0.066$ |
| $K = 10$ | $\mathbf{0.791 \pm 0.061}$ |

from very few references. The conservative thresholds result in high precision at the cost of moderate recall, reflecting an explicit design choice: in PAM applications, false positives for rare or novel species carry higher ecological cost than missed detections. This performance requires no model retraining: novel species are integrated by computing prototype embeddings from the support set.

### E. COMPONENT ABLATION

We quantify the role of each component under a common reduced-budget training harness (Table 6); absolute values are therefore below the full-protocol results of Table 2 and the table should be read for relative comparison. Progressive unfreezing of the backbone has the clearest effect; the combined loss preserves prototype accuracy while

**TABLE 6. Component ablation under a common reduced-budget harness (relative comparison; absolute values are below table 2).**

| Configuration | Proto-acc. | Internal AUROC | Separation |
|---|---|---|---|
| ArcFace only | 0.885 | 0.876 | 1.53 |
| Triplet only | 0.856 | 0.894 | 1.98 |
| Combined (full) | 0.887 | 0.879 | 1.54 |
| Combined + random sampling | 0.893 | 0.902 | 1.52 |
| Combined + frozen backbone | 0.863 | 0.865 | 1.38 |

**TABLE 7. Model size and single-GPU inference cost.**

| Quantity | Value |
|---|---|
| Parameters | 86.7 M |
| Model size (fp32) | 347 MB |
| Latency (GPU) | 18.3 ms/clip |
| Throughput (GPU) | 54.6 clips/s |
| Prototype DB | 10.2 MB / 10k clips |

increasing inter-class separation relative to ArcFace alone, whereas triplet alone trades accuracy for separation. Balanced and random class sampling perform comparably; we adopt balanced (P-K) sampling because it guarantees positive pairs per class for the triplet term—a training-stability consideration rather than an accuracy gain.

### F. COMPUTATIONAL COST

The model (the AST backbone plus a lightweight projection head, 86.7 M parameters, 347 MB in fp32) runs inference at 54.6 clips/s (18.3 ms per clip) on a single GPU, well within near-real-time monitoring requirements (Table 7). The embedding-and-prototype design adds new species at deployment time without retraining, at a storage cost of about 10 MB per 10k reference clips.

## VI. DISCUSSION

### A. ADVANTAGES OF METRIC LEARNING FOR MARINE BIOACOUSTICS

The metric learning framework offers several advantages for passive acoustic monitoring (PAM). Unlike traditional closed-set classifiers, this approach inherently enables open-set recognition; by learning a discriminative embedding space rather than a fixed mapping to class labels, the model can identify out-of-distribution signals—such as unknown species as points that lie at a significant distance from all known clusters.

Moreover, the structured embedding space enables few-shot inference without the need for model retraining. Novel species can be integrated into the system by simply providing a limited support set of reference embeddings, allowing the architecture to adapt to new contexts in real-time. This geometric formulation also provides a natural handling of class imbalance, often an issue in bioacoustics where rare species can be under-represented; by focusing on relative distances rather than global class frequencies, the model maintains high discriminative power even for sparsely sampled categories.

**TABLE 8. Holdout species confusion: nearest known prototype and cosine similarity.**

| Holdout | Nearest Known | Sim. |
|---|---|---|
| Bearded Seal | Bowhead Whale | 0.55 |
| Bottlenose Dolphin | False Killer Whale | 0.45 |
| Leopard Seal | Humpback Whale | 0.63 |
| Narwhal | Fraser's Dolphin | 0.37 |
| Ross Seal | Killer Whale | 0.47 |
| Walrus | Sperm Whale | 0.60 |

From a practical perspective, the resulting interpretable embedding space - where Euclidean or cosine distances directly correlate with similarity - may offer researchers a useful tool for inspecting model decisions. For PAM systems deployed in remote or non-stationary environments, these features support extensibility to new species without retraining.

### B. COMPARISON WITH CLOSED-SET APPROACHES

The proposed metric embedding framework offers several potential advantages over traditional closed-set architectures. In standard cross-entropy-based models, the network learns to partition the feature space into a fixed number of regions corresponding to the training classes. While effective for static datasets, this approach lacks the flexibility required for dynamic marine environments.

Metric learning is well-suited when dealing with a large number of species and limited samples per class. By optimizing the angular margin between embeddings (e.g., via Cosine Similarity), the model focuses on learning the intrinsic acoustic "fingerprint" of a species rather than simply finding a decision boundary. This can lead to feature representations that transfer to species not seen during training, as shown in our few-shot evaluation. In our tests, the AST-based embedding demonstrated discriminative power in few-shot scenarios, where traditional Softmax layers would require an architectural change for the output layer and extensive retraining to incorporate even a single new species.

Our approach, by contrast, naturally supports open-set recognition: by measuring the distance to the nearest neighbor in the embedding space, the system can flag vocalizations that fall outside a predefined threshold as "unknown" or "potential new species." This allows for a potentially scalable deployment where new reference signatures can be added to the gallery database without ever modifying the underlying neural network weights.

A key advantage of the proposed framework lies in the decoupling of the feature extractor (the AST backbone) from the classification logic, which is managed through a Dynamic Reference Database (also referred to as a gallery). Unlike traditional closed-set classifiers our approach treats the neural network as a universal acoustic encoder that maps sounds into a structured vector space.

The user can update the model's knowledge "on-the-fly" simply by adding a small set of validated recordings to the Reference Database. The AST model automatically

transforms these new samples into embeddings, which serve as new anchors for the k-Nearest Neighbor (k-NN) search.

This means the monitoring system can be specialized for a specific geographic region or task in seconds by simply populating the database with local species of interest. As demonstrated by our Few-Shot tests in Table 5, the discriminative power of the embedding space is often robust enough to allow for high-accuracy recognition even with a minimal number of support samples. This significantly reduces the need for large annotated datasets in new deployment scenarios.

A small drawback is that metric learning introduces a slightly more complex training pipeline and more epochs to converge, and it requires careful triplet selection.

### C. LIMITATIONS AND FUTURE WORK

Our closed-set accuracy does not exceed a softmax baseline, and on pure few-shot transfer a frozen bioacoustic model (AVES) is competitive; we do not regard these as weaknesses, since the value of the system lies in delivering closed-set recognition, open-set detection and few-shot discovery from a single model under a leakage-free protocol—a combination none of the baselines provides. The competitiveness of AVES suggests that large-scale bioacoustic pre-training and our task-specific supervision are complementary, motivating future fine-tuning of such encoders with our objective.

Several further limitations should be acknowledged. First, evaluation uses a single curated database; cross-dataset experiments on continuous PAM streams would strengthen generalization claims.

We note that several properties of our evaluation are deliberate design choices rather than oversights. The Watkins database is curated and relatively clean, the audio is resampled to a common 16 kHz, and clips are a fixed 10 s (160,000 samples); this controlled setting isolates the question we set out to answer—whether a single embedding can support closed-set, open-set and few-shot recognition jointly—without confounding it with the segmentation and denoising challenges of raw field recordings. Establishing the behavior of the method under these conditions is a prerequisite for, not a substitute for, deployment on continuous streams, which we treat as the next step rather than a limitation of the present design.

We have not evaluated the framework on overlapping vocalizations or high shipping-noise conditions typical of continuous PAM deployments; under such conditions we would expect degraded separation between query embeddings and prototypes, likely increasing the false-positive rate of the open-set rejection step. Robustness to these field conditions is a natural target for the detection-front-end integration discussed above.

Second, the limited recording diversity per species (2–7 sessions) constrains evaluation rigor. We adopted clip-level splitting consistent with the literature, acknowledging possible weak temporal correlation. The open-set and few-shot evaluations on held-out species are immune to this concern.

Third, the AST backbone emphasizes the 50 Hz–8 kHz range; mysticete calls below 100 Hz fall near the edge of this range. Bioacoustic foundation models represent promising backbone alternatives.

Fourth, the framework assumes pre-segmented vocalizations; integration with detection front-ends is needed for end-to-end deployment.

We see a natural three-phase path to operational use. *Near term*, the embedding can be coupled to an existing detection front-end and run on continuous PAM streams, turning the pre-segmented evaluation here into an end-to-end pipeline; this is primarily an integration effort, since inference already runs in near-real time on a single GPU (Table 7). *Medium term*, the prototype store enables on-site enrollment of new species and individuals from a few reference clips, with active-learning selection driven by embedding geometry to prioritize the most informative recordings for annotation. *Longer term*, distilling or quantizing the backbone for edge deployment on autonomous recorders would bring the storage and compute footprint—currently modest but GPU-oriented (Table 7)—within the budget of battery-powered field hardware, and pairing the embedding with generative augmentation could address the most data-poor species. Domain adaptation to operational datasets and bioacoustic-specific backbones cut across all three phases.

## VII. CONCLUSION

This work introduced a metric embedding framework for marine mammal acoustic classification, combining an AudioSet-pretrained AST backbone with a 256-dimensional projection head trained under ArcFace and online triplet loss. The framework projects vocalizations into a hyperspherical embedding space where cosine distance encodes acoustic similarity.

1 Evaluation on the Watkins database across 25 known and 6 held-out species, averaged over three independent runs, demonstrates: closed-set classification at $92.4 \pm 0.5\%$ accuracy (0.926 macro F1, on par with a softmax baseline), open-set detection at AUROC 0.891 (leave-$N$-species-out), and few-shot discovery at macro F1 0.791 ($K = 10$)—all from the same model without retraining. The embedding geometry (separation ratio $5.83\times$) produces interpretable similarity relationships aligned with acoustic and phylogenetic structure.

These results support metric embedding learning as a promising paradigm for marine mammal acoustic monitoring, with potential to accommodate the open-ended nature of real-world PAM deployments.

By supporting rejection of unknown queries and few-shot integration of novel species, the framework points toward scalable, adaptive marine bioacoustic monitoring systems.

*Reproducibility.* All models were re-trained from scratch with fixed seeds (42, 43, 44) under a single verified configuration (angular margin 0.35, triplet weight 1.1,

**TABLE 9.** Per-species closed-set performance (representative run, seed 42), ordered by test-set size *n*. CIs are 95% bootstrap intervals on F1.

| Species | n | Prec. | Rec. | F1 | 95% CI |
|---|---|---|---|---|---|
| Spinner Dolphin | 15 | 0.882 | 1.000 | 0.938 | [0.86, 1.00] |
| Striped Dolphin | 11 | 1.000 | 1.000 | 1.000 | [1.00, 1.00] |
| Frasers Dolphin | 11 | 0.909 | 0.909 | 0.909 | [0.76, 1.00] |
| Humpback Whale | 9 | 1.000 | 1.000 | 1.000 | [1.00, 1.00] |
| Rissos Dolphin | 9 | 1.000 | 1.000 | 1.000 | [1.00, 1.00] |
| Short-Finned Pacific Pilot Whale | 9 | 1.000 | 1.000 | 1.000 | [1.00, 1.00] |
| Long-Finned Pilot Whale | 9 | 1.000 | 0.889 | 0.941 | [0.80, 1.00] |
| Pantropical Spotted Dolphin | 9 | 0.875 | 0.778 | 0.824 | [0.62, 1.00] |
| Sperm Whale | 9 | 0.857 | 0.667 | 0.750 | [0.46, 0.94] |
| False Killer Whale | 8 | 1.000 | 1.000 | 1.000 | [1.00, 1.00] |
| Atlantic Spotted Dolphin | 8 | 0.889 | 1.000 | 0.941 | [0.84, 1.00] |
| Bowhead Whale | 8 | 1.000 | 0.875 | 0.933 | [0.77, 1.00] |
| White-sided Dolphin | 8 | 1.000 | 0.875 | 0.933 | [0.77, 1.00] |
| Clymene Dolphin | 8 | 0.875 | 0.875 | 0.875 | [0.67, 1.00] |
| White-beaked Dolphin | 8 | 1.000 | 0.750 | 0.857 | [0.55, 1.00] |
| Melon Headed Whale | 8 | 0.667 | 1.000 | 0.800 | [0.70, 0.94] |
| Common Dolphin | 7 | 1.000 | 1.000 | 1.000 | [1.00, 1.00] |
| Northern Right Whale | 7 | 1.000 | 1.000 | 1.000 | [1.00, 1.00] |
| Harp Seal | 7 | 0.875 | 1.000 | 0.933 | [0.82, 1.00] |
| Beluga | 6 | 1.000 | 1.000 | 1.000 | [1.00, 1.00] |
| Fin Finback Whale | 6 | 1.000 | 1.000 | 1.000 | [1.00, 1.00] |
| Rough-Toothed Dolphin | 6 | 0.750 | 1.000 | 0.857 | [0.71, 1.00] |
| Killer Whale | 5 | 1.000 | 0.600 | 0.750 | [0.33, 1.00] |
| Minke Whale | 3 | 1.000 | 1.000 | 1.000 | [1.00, 1.00] |
| Southern Right Whale | 3 | 1.000 | 1.000 | 1.000 | [1.00, 1.00] |
| **Macro Average** | **197** | **0.943** | **0.929** | **0.930** | — |

batch size 16, learning rate $2 \times 10^{-4}$, three warm-up epochs); the released code, the hyperparameter table and the trained checkpoints are mutually consistent. Because every reported value is regenerated from this retraining and the corrected validation-only threshold protocol, absolute numbers differ slightly from the original submission while the conclusions are unchanged.

## ACKNOWLEDGMENT

The authors acknowledge the Watkins Marine Mammal Sound Database which is maintained by the Woods Hole Oceanographic Institution and the New Bedford Whaling Museum, for providing the acoustic data used in this study. They also acknowledge the use of Claude (Anthropic) for language proofreading assistance.

## APPENDIX

## PER-SPECIES CLOSED-SET RESULTS

Table 9 reports per-species precision, recall and F1 on the test set for the representative run (seed 42, macro F1 0.930), with 95% bootstrap confidence intervals on F1. Species are ordered by the number of test clips $n$. Several species reach an F1 of 1.000, but for those with few test clips ($n < 8$, lower part of the table) the confidence intervals are wide and these point estimates should be interpreted with caution; the three-seed macro F1 reported in the main text ($0.926 \pm 0.005$, Table 2) is the more reliable aggregate. Note that when every test clip of a species is classified correctly, the bootstrap resampling distribution is concentrated at F1=1.0 and the interval collapses to [1.00, 1.00] (e.g., Minke Whale and Southern Right Whale, $n$=3); such degenerate intervals reflect the absence of observed errors in a very small sample rather than genuine certainty, and should be read together with $n$.

## REFERENCES

[1] D. Duan, L.-G. Lü, Y. Jiang, Z. Liu, C. Yang, J. Guo, and X. Wang, "Real-time identification of marine mammal calls based on convolutional neural networks," *Appl. Acoust.*, vol. 192, Apr. 2022, Art. no. 108755.

[2] S. Nanaware, R. Shastri, Y. Joshi, and A. Das, "Passive acoustic detection and classification of marine mammal vocalizations," in *Proc. Int. Conf. Commun. Signal Process.*, Apr. 2014, pp. 493–497.

[3] P. J. Clemins, M. B. Trawicki, K. Adi, J. Tao, and M. T. Johnson, "Generalized perceptual features for vocalization analysis across multiple species," in *Proc. IEEE Int. Conf. Acoust. Speed Signal Process.*, 2006, pp. 253–256.

[4] A. K. Ibrahim, H. Zhuang, L. M. Chérubin, N. Erdol, G. O'corry-Crowe, and A. M. Ali, "A multimodel deep learning algorithm to detect North Atlantic right whale up-calls," *J. Acoust. Soc. Amer.*, vol. 150, no. 2, pp. 1264–1272, Aug. 2021.

[5] M. Zhong and W. Cai, "Marine mammal sound recognition based on feature fusion," *Electron. Sci. & Tech.*, vol. 32, no. 5, pp. 32–37, 2019.

[6] W. Cai, J. Zhu, M. Zhang, and Y. Yang, "A parallel classification model for marine mammal sounds based on multi-dimensional feature extraction and data augmentation," *Sensors*, vol. 22, no. 19, p. 7443, Sep. 2022.

[7] Y. Shiu, K. J. Palmer, M. A. Roch, E. Fleishman, X. Liu, E.-M. Nosal, T. Helble, D. Cholewiak, D. Gillespie, and H. Klinck, "Deep neural networks for automated detection of marine mammal species," *Sci. Rep.*, vol. 10, no. 1, p. 607, Jan. 2020.

[8] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 776–780.

[9] E. L. White, P. R. White, J. M. Bull, D. Risch, S. Beck, and E. W. J. Edwards, "More than a whistle: Automated detection of marine sound sources with a convolutional neural network," *Frontiers Mar. Sci.*, vol. 9, Oct. 2022, Art. no. 879145.

[10] W. Cheng, H. Chen, J. Jiang, S. Li, J. Wang, and Y. Zhou, "Recognition and classification techniques of marine mammal calls based on LSTM and expanded causal convolution," *Frontiers Mar. Sci.*, vol. 12, May 2025, Art. no. 1603090.
[11] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.
[12] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4685–4694.
[13] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. E. Boult, "Toward open set recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 7, pp. 1757–1772, Jul. 2013.
[14] S. Bouma, M. D. M. Pawley, K. Hupman, and A. Gilman, "Individual common dolphin identification via metric embedding learning," in *Proc. Int. Conf. Image Vis. Comput. New Zealand (IVCNZ)*, Nov. 2018, pp. 1–6.
[15] J. Liang, I. Nolasco, B. Ghani, H. Phan, E. Benetos, and D. Stowell, "Mind the domain gap: A systematic analysis on bioacoustic sound event detection," in *Proc. 32nd Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2024, pp. 1257–1261.
[16] I. Moummad, N. Farrugia, and R. Serizel, "Regularized contrastive pre-training for few-shot bioacoustic sound detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2024, pp. 1436–1440.
[17] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio spectrogram transformer," in *Proc. Interspeech*, Aug. 2021, pp. 571–575.
[18] D. Gillespie, D. K. Mellinger, J. Gordon, D. McLaren, P. Redmond, R. McHugh, P. Trinder, X.-Y. Deng, and A. Thode, "Pamguard: Semiautomated, open source software for real-time acoustic detection and localization of cetaceans," *J. Acoust. Soc. Amer.*, vol. 125, no. 4, p. 2547, 2009.
[19] J. Wang, S. Jiang, M. Wang, W. Shi, L. Xu, and S. Yan, "A deep learning model for marine mammal call classification," *IEEE J. Ocean. Eng.*, vol. 50, no. 4, pp. 2642–2660, Oct. 2025.
[20] D. T. Murphy, E. Ioup, M. T. Hoque, and M. Abdelguerfi, "Residual learning for marine mammal classification," *IEEE Access*, vol. 10, pp. 118409–118418, 2022.
[21] X. Liu, X. Liu, S. Du, and J. Cheng, "Hear you say you: An efficient framework for marine mammal sounds' classification," in *Proc. AAAI Conf. Artif. Intell.*, 2024, vol. 38, no. 20, pp. 22250–22257.
[22] Q. Yi, C. Xie, D. Guan, and W. Yuan, "Man2Marine: Marine mammal sound classification in small samples by transfer learning from human sound data," *Pattern Recognit. Lett.*, vol. 190, pp. 185–191, Apr. 2025.
[23] A. Kolesnikov, A. Dosovitskiy, D. Weissenborn, G. Heigold, J. Uszkoreit, L. Beyer, M. Minderer, M. Dehghani, N. Houlsby, S. Gelly, T. Unterthiner, and X. Zhai, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 1–5.
[24] R. Schwinger, P. Vali Zadeh, L. Rauch, M. Kurz, T. Hauschild, S. Lapp, and S. Tomforde, "Foundation models for bioacoustics - a comparative review," 2025, *arXiv:2508.01277*.
[25] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.
[26] M. Hagiwara, "AVES: Animal vocalization encoder based on self-supervision," 2022, *arXiv:2210.14493*.
[27] S. Kahl, C. M. Wood, M. Eibl, and H. Klinck, "BirdNET: A deep learning solution for avian diversity monitoring," *Ecol. Informat.*, vol. 61, Mar. 2021, Art. no. 101236.
[28] B. Ghani, T. Denton, S. Kahl, and H. Klinck, "Global birdsong embeddings enable superior transfer learning for bioacoustic classification," *Sci. Rep.*, vol. 13, no. 1, p. 22876, Dec. 2023.
[29] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, and F. Wei, "BEATs: Audio pre-training with acoustic tokenizers," 2022, *arXiv:2212.09058*.
[30] K. Musgrave, S. Belongie, and S.-N. Lim, "A metric learning reality check," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 681–699.
[31] (2023). *Watkins Marine Mammal Sound Database*. [Online]. Available: https://whoicf2.whoi.edu/science/B/whalesounds/
[32] Q. Lhoest et al., "Datasets: A community library for natural language processing," in *Proc. Conf. Empirical Methods Natural Lang. Process., Syst. Demonstrations*, 2021, pp. 175–184.
[33] M. Hagiwara, B. Hoffman, J.-Y. Liu, M. Cusimano, F. Effenberger, and K. Zacarian, "BEANS: The benchmark of animal sounds," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.
[34] D. Robinson, M. Miron, M. Hagiwara, B. Weck, S. Keen, M. Alizadeh, G. Narula, M. Geist, and O. Pietquin, "NatureLM-audio: An audio-language foundation model for bioacoustics," 2024, *arXiv:2411.07186*.
[35] R. J. Tibshirani and B. Efron, "An introduction to the bootstrap," *Monographs Statist. Appl. Probab.*, vol. 57, no. 1, pp. 1–436, 1993.
[36] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach," *Biometrics*, vol. 44, no. 3, pp. 837–845, Sep. 1988.
[37] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.
[38] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 18661–18673.

**VINCENZO LIPARI** (Member, IEEE) received the M.Sc. degree in telecommunication engineering and the Ph.D. degree in information technology from the Politecnico di Milano, Milan, Italy, in 2002 and 2006, respectively.

From 2016 to 2023, he was a Postdoctoral Researcher with the Geophysical Prospecting Group and the Image and Sound Processing Laboratory (ISPL), Politecnico di Milano. Since 2023, he has been a Senior Researcher with the National Institute of Oceanography and Geophysics (OGS), Trieste, Italy. His research interests include signal processing, machine learning, large-scale inverse problems, seismic processing, seismic imaging, and signal processing in general. He is also a gold member of EAGE and an Active Member of SEG. He is an Associate Editor of the *Bulletin of Geophysics and Oceanography*. He is also a Referee for several journals, including IEEE Transactions on Geoscience and Remote Sensing and IEEE Geoscience and Remote Sensing Letters and Geophysics.

**GIUSEPPE BRANCATELLI** received the M.Sc. degree in environmental engineering and the Ph.D. degree in geophysics from the University of Trieste, Italy, in 2006 and 2010, respectively.

From 2011 to 2013, he was a Postdoctoral Researcher with the Department of Civil and Environmental Engineering, University of Trieste. Since 2015, he has been a Researcher with the National Institute of Oceanography and Applied Geophysics (OGS), Trieste, Italy. His research interests include seismic data processing, velocity modeling, and depth imaging.

**EDY FORLIN** received the M.Sc. degree in geological sciences and the Ph.D. degree in environmental sciences, with a focus on the physical marine and coastal environment from the University of Trieste, Trieste, Italy, in 1998.

Since 2011, he has been a Technologist with the National Institute of Oceanography and Applied Geophysics, Trieste. Previously, he was a Post-doctoral Researcher with the University of Trieste and a Marine Geophysicist in the private sector. His work focuses on applied geophysics, seismic data processing and interpretation, and marine geophysical surveys. His main research interests include 2D seismic processing, the reprocessing of vintage seismic data, time and depth seismic imaging, morpho-bathymetric data analysis, seismic and geological hazard assessment, and the geophysical characterization of sites for geological $CO_2$ storage. He has contributed to several national and international research and service projects, often with technical or scientific responsibility. He has also been involved in teaching, training, and mentoring activities in seismic data processing and applied geophysics, including the supervision of graduate students and internships.