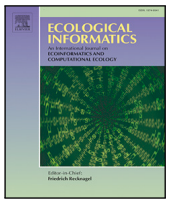
Ecological Informatics 96 (2026) 103816

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf



A noise-masking marine mammal sound classification method via multi-task learning

Xuchen Wang, Yougan Chen *, Kunyun Du, Yihao Zhao, Yanhan Dong, Shen'ao Tu, Yi Tao, Xiaomei Xu

College of Ocean and Earth Sciences, Xiamen University, Xiamen, 361102, China
Shenzhen Research Institute of Xiamen University, Shenzhen, 518000, China
Fujian Ocean Innovation Center, Xiamen, 361102, China
Key Laboratory of Underwater Acoustic Communication and Marine Information Technology (Xiamen University), Ministry of Education, Xiamen, 361005, China

ARTICLE INFO

Dataset link: https://github.com/XC4Marine/MT-MaskNet

Keywords:
Marine mammals recognition
Passive acoustic monitoring
Multi-task learning
Gating mechanism
Noise-masking

ABSTRACT

Underwater noise pollution hinders passive acoustic monitoring (PAM) of marine mammals, as ambient noise masks target signals and degrades classification performance. To address this, we propose MT-MaskNet, an end-to-end multi-task learning (MTL) framework based on ResNet18. This framework jointly optimizes marine mammal sound classification and noise-type classification as the primary and auxiliary tasks, respectively. Guided by the auxiliary branch, a mid-level gating mechanism dynamically suppresses noise-related activations while preserving salient acoustic patterns. We adopt a two-stage training strategy to decouple sound and noise representations before fine-tuning the gating module. This approach mitigates negative transfer and promotes positive transfer to the primary task. On a synthesized dataset featuring three dolphin species with realistic noise superposition, MT-MaskNet achieved a mean accuracy of 95.00% ± 0.8%. This performance significantly outperformed single-task baselines ($p = 0.0018$). Evaluations on real-world PAM recordings further demonstrated a marked accuracy improvement from 44.00% to 73.75%. Overall, the gating mechanism enables robust feature enhancement through coarse-grained noise indication with minimal computational overhead.

1. Introduction

Marine mammals are essential for sustaining the structural integrity and equilibrium of marine ecosystems. However, intensifying anthropogenic activities — including seismic surveys, offshore wind farms, and commercial shipping — severely threaten marine mammal populations and their habitats through noise pollution and environmental degradation. Effective conservation necessitates advanced technologies for real-time recognition of marine mammals, enabling stakeholders to redirect human activities away from sensitive habitats and migration corridors, thereby minimizing disturbance to protected species (Cai et al., 2022). Moreover, reliable species recognition techniques also enables assessment of population distribution and identification of adverse environmental factors, supporting biodiversity preservation and ecosystem management.

The diverse sounds produced by marine organisms form the basis for Passive Acoustic Monitoring (PAM) and automated sound-based species recognition. Unlike active sonar, PAM is non-invasive, as it does not emit signals and thus causes negligible interference with the marine environment (Cauchy et al., 2023). Early approaches to marine mammal sound classification relied on hand-crafted features derived from time-domain, frequency-domain, cepstral, and statistical analyses. Subsequent statistical learning methods shifted toward data-driven recognition, which substantially reduced the labor required for manual annotation (Shiu et al., 2020). Nevertheless, these feature-engineering-dependent approaches often lack generalization across datasets collected at different sampling rates, geographic regions, or recording platforms.

PAM datasets have expanded rapidly in volume and diversity, driven by the declining costs of acoustic data acquisition and storage. While traditional statistical models perform well on small, controlled datasets, their accuracy declines on large-scale, heterogeneous PAM data. To address these shortcomings, deep learning (DL) has demonstrated superior feature representation and generalization capabilities in numerous domains, prompting extensive exploration for marine mammal recognition (Aslam et al., 2024). In particular, convolutional neural networks (CNNs) have outperformed conventional machine learning methods in bio-acoustic classification tasks owing to their

* Corresponding author.
E-mail address: chenyougan@xmu.edu.cn (Y. Chen).

https://doi.org/10.1016/j.ecoinf.2026.103816
Received 26 October 2025; Received in revised form 8 May 2026; Accepted 9 May 2026
Available online 15 May 2026

exceptional ability to extract local features from image-like representations. Among these, ResNet18 (He et al., 2015) has proven especially effective, addressing overfitting and vanishing gradient problems through residual connections and achieving strong performance in both image and speech recognition domains. Typical inputs for CNN-based marine mammal sound classifiers include acoustic spectrograms (Shiu et al., 2020; White et al., 2022), Mel spectrograms, and Mel-frequency cepstral coefficients (MFCCs) (Li, 2023), which undergo automatic extraction before being fed into deep models for end-to-end training (Lei et al., 2022). Collectively, these studies demonstrate the effectiveness of CNN architectures for marine mammal sound recognition.

In recent years, increasing attention has shifted toward Transformer-based architectures. Unlike CNN classifiers, Transformers leverage self-attention mechanisms that excel at learning contextual relationships and global modeling in bio-acoustic tasks with inherent temporal structures (Atito et al., 2024; Schäfer-Zimmermann et al., 2026). In marine mammal sound recognition, the application of Transformers remains in an exploratory phase. For instance, Cotillard et al. (2024) demonstrated that Transformers achieved higher classification accuracy for beluga whale calls. Some studies have also explored pre-training on large-scale general audio datasets followed by fine-tuning for marine mammal tasks (Zhang et al., 2025). Nevertheless, fine-tuning such models typically incurs higher memory consumption and computational demands than training a purpose-built compact CNN from scratch.

Various underwater ambient noises from anthropogenic activities and natural sources mask marine mammal sound, presenting a major obstacle to accurate species identification and robust classification. Existing masking techniques, such as Ideal Binary Mask (IBM) (Olatinwo and Seto, 2025) and pseudo-attention masks (Razig et al., 2025), aim to suppress noise while enhancing salient acoustic components. However, because these approaches typically operate as discrete preprocessing steps, they incur additional computational overhead and risk early-stage information loss. Moreover, attention-based feature-level masking, while powerful, often entails relatively high computational complexity during both training and inference (Vaswani et al., 2017).

To overcome these limitations, we propose a multi-task learning (MTL) framework (Crawshaw, 2020) with ResNet18 as its backbone for marine mammal sound classification. By concurrently training on species classification (primary task) and noise-type identification (auxiliary task), the model extracts explicit and complementary feature representations. Critically, we introduce a gating mechanism at the intermediate feature level, guided by the auxiliary noise branch, to dynamically suppress noise-related activations while preserving and enhancing target acoustic features. This end-to-end architecture eliminates the need for separate denoising modules and enables efficient feature sharing without information loss.

In practical PAM scenarios, noise-type labels are often readily accessible from field deployments, such as monitoring programs during offshore wind farm construction (e.g., SERCEL QUIETSEA system, Yetra Tech Neptune AI project) or targeted studies in high-interference environments (Vishnu et al., 2024). Leveraging such labeled data, the proposed MTL-based masking strategy equips the model to extract multi-faceted features from a single input and better disentangle overlapping noise-signal patterns, thereby improving robustness in real-world underwater acoustic environments. The main contributions of this work are summarized as follows:

(1) Proposes an end-to-end multi-task framework that integrates noise-type classification as an auxiliary task with a mid-level gating mechanism, enabling adaptive feature-level suppression of diverse real-world underwater noise.

(2) Introduces a two-stage training strategy that first decouples sound and noise representations and subsequently fine-tunes the gating module, effectively mitigating negative transfer while maximizing positive transfer to the primary sound classification task.

(3) Demonstrates, through rigorous evaluation on both synthesized and real-world PAM data, that the gating primarily provides coarse-grained noise indication to enhance main-task robustness, achieving substantial performance gains and offering mechanistic insights into MTL for bio-acoustic applications.

The remainder of the paper is organized as follows: Section 2 reviews existing related research; Section 3 elaborates on the methods used. Section 4 presents the experimental results and analysis. Section 5 discusses the performance, limitations and future work. Section 6 concludes the paper's research.

2. Related work

CNNs have demonstrated potential in classifying marine mammal sounds within PAM datasets, outperforming traditional methods in generalization and adaptability (Shiu et al., 2020). Recent supervised research on marine mammal sound classification has enhanced feature representation by integrating multi-channel inputs (White et al., 2022), multi-type feature integration (Li, 2023), multi-granularity extraction (Li et al., 2024), and multi-scale feature combination (Hamard et al., 2024). These strategies have proven effective in improving robustness and decision-making.

More recently, MTL has streamlined underwater sound classification by sharing representations across task branches (Crawshaw, 2020). This architectural synergy enhances data efficiency and alleviates the reliance on massive labeled datasets. In practice, Huang et al. (2025) employed a shared feature extractor to jointly optimize classification and reconstruction tasks, markedly improving few-shot performance in marine mammal recognition. Similarly, Li (2023) optimized signal recovery, frequency selection, and classification concurrently to extract more robust features for underwater target recognition. In summary, MTL provides an effective framework for underwater bio-acoustic tasks by promoting feature sharing, mitigating data scarcity issues, and enabling more robust classification under marine environments.

Despite the advantages of MTL, supervised models still depend heavily on abundant high-quality labeled data to achieve robust generalization. In real marine environments, however, PAM recordings are frequently corrupted by diverse ambient noise, resulting in signal masking and scarce well-annotated samples. To isolate target sounds from noise interference, researchers have employed noise masking techniques, such as the IBM, as a preprocessing step (Olatinwo and Seto, 2024) step to enhance upcall detection of North Atlantic right whales in low-SNR (Signal-to-Noise Ratio) conditions. Similarly, Razig et al. (2025) utilized Gaussian soft masks to direct model attention toward biologically relevant acoustic regions in estuarine environments. However, these methods often require additional frontend denoising modules. As an alternative, attention mechanisms can implicitly suppress noise by focusing on salient signal regions. Deng and Hong (2025) employed attention to attenuate environmental noise and amplify gradient responses to target signals in low-SNR scenarios. Nevertheless, attention-based methods often suffer from high computational complexity in both time and space. In contrast, gating mechanisms offer a more efficient alternative by controlling information flow through learnable gates. When guided by noise-type labels, gating mechanism offers lower computational overhead (Gu et al., 2020) and carries more explicit physical interpretability. With the growing body of research on PAM under specific noise conditions (Vishnu et al., 2024), noise-type labels are becoming increasingly available and reliably recorded, providing a practical foundation for label-guided gating strategies.

The proposed MTL framework is built on ResNet18. It treats marine mammal sound classification as the primary task and noise-type classification as the auxiliary task. We introduce a lightweight gating mechanism at intermediate feature layers. Guided by readily available noise labels, the model dynamically suppresses noise-related activations while preserving and enhancing salient acoustic features of target sounds. This design delivers a robust and computationally efficient solution for marine mammal sound classification in complex, heterogeneous PAM environments.

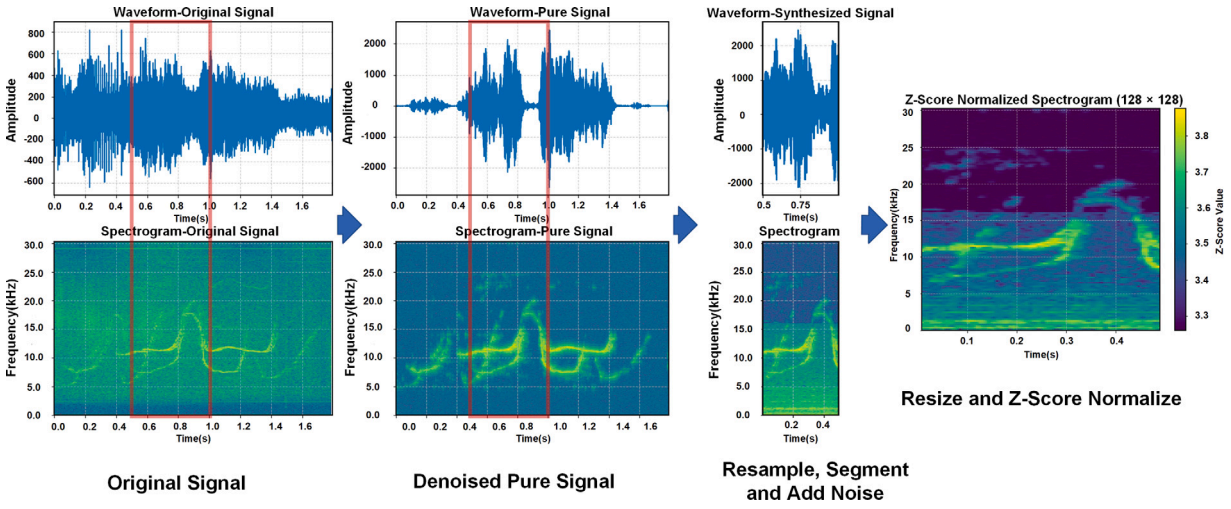


Fig. 1. Data pre-process pipeline.

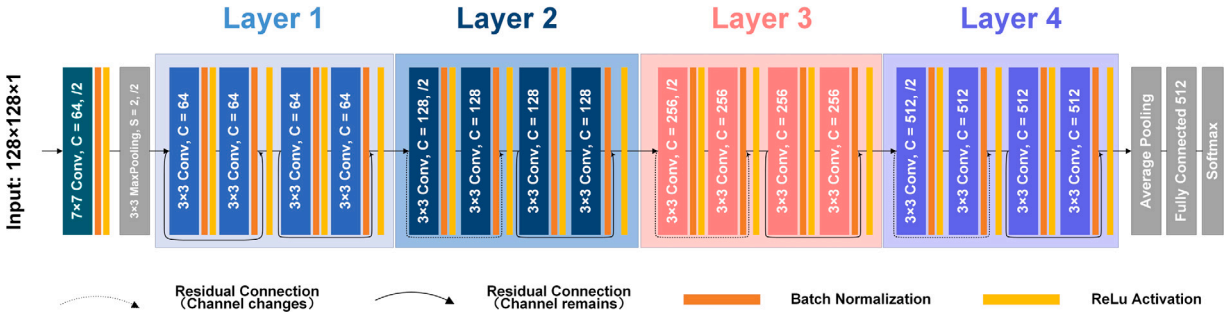


Fig. 2. ResNet18 architecture.

3. Methodology

3.1. Data preprocessing

The overall preprocessing pipeline is illustrated in Fig. 1. We downsampled all acoustic recordings to 60.6 kHz, matching the lowest native sampling rate in the original dataset. To mitigate data scarcity and improve generalization, we applied data augmentation by superimposing realistic ambient noise onto relatively pure target signals, following the approach of Nanni et al. (2020). To generate high-quality training pairs, we performed Singular Value Decomposition (SVD) denoising by retaining only the top-20 singular values (Zhang, 2015). Notably, we restrict this step to dataset preparation; the model processes raw field recordings during inference to maintain real-world applicability. The processed signals were segmented into 0.5 s clips with zero-padding applied to shorter segments to maintain temporal uniformity. This duration suffices to capture transient clicks and the contextual frequency-modulated patterns of whistles, consistent with previous studies that utilized 0.5 s spectrograms for odontocete click detection (Bermant et al., 2019) and aligns with typical odontocete call durations of 0.5–1.5 s (Palmero et al., 2023).

Four representative marine ambient noise types were selected — wind&wave, rain, vessel, and snapping shrimp — as they collectively occupy most of the frequency bands relevant to marine mammal sounds. Each clean signal was mixed with these noise types at five controlled SNR levels: 0 dB, 5 dB, 10 dB, 15 dB, and 20 dB. These SNR levels were chosen to span a practically relevant and challenging range encountered in real underwater PAM scenarios, thereby exposing the model to a diverse set of interference conditions during training and promoting improved robustness and adaptability.

Marine mammal sounds are typically non-stationary signals characterized by time-varying, and complex modulation patterns. In this study, Short-Time Fourier Transform (STFT) is applied to generate log-scaled spectrograms, which serve as the primary input features for deep learning models. The resulting spectrograms are resampled to a uniform resolution of 128 × 128 pixels using bilinear interpolation (Lü et al., 2024). Finally, Z-score normalization is performed on the decibel-scaled spectrograms to remove absolute amplitude variations while preserving relative structural information (Fei et al., 2021; Peng et al., 2024). To prevent data leakage and ensure fair evaluation, normalization parameters for the validation and test sets are computed exclusively from the training set statistics.

3.2. Backbone architecture: ResNet18

We utilize ResNet18 (He et al., 2015) as the architectural backbone to mitigate overfitting and address the vanishing gradient problem common in deep networks. ResNet18 comprises 18 learnable layers, including an initial 7 × 7 convolutional layer followed by four stages of residual blocks. Each residual block consists of two 3 × 3 convolutional layers with batch normalization and ReLU activation, connected via an identity shortcut. When dimensions mismatch, a 1 × 1 convolution is used for projection in the shortcut path. Fig. 2 illustrates the overall architecture. Formally, the output of a residual block is defined as:

$$y = ReLu(F(x, W_i)) + x \tag{1}$$

where $F(x, W_i)$ denotes the residual function learned by the stacked layers, and x is the input to the block. This formulation ensures that gradients can propagate directly through the identity path during back-propagation:

$$\frac{\partial L}{\partial x} = \frac{\partial L}{\partial y}\left(\frac{\partial F(x, W_i)}{\partial x} + I\right) \tag{2}$$

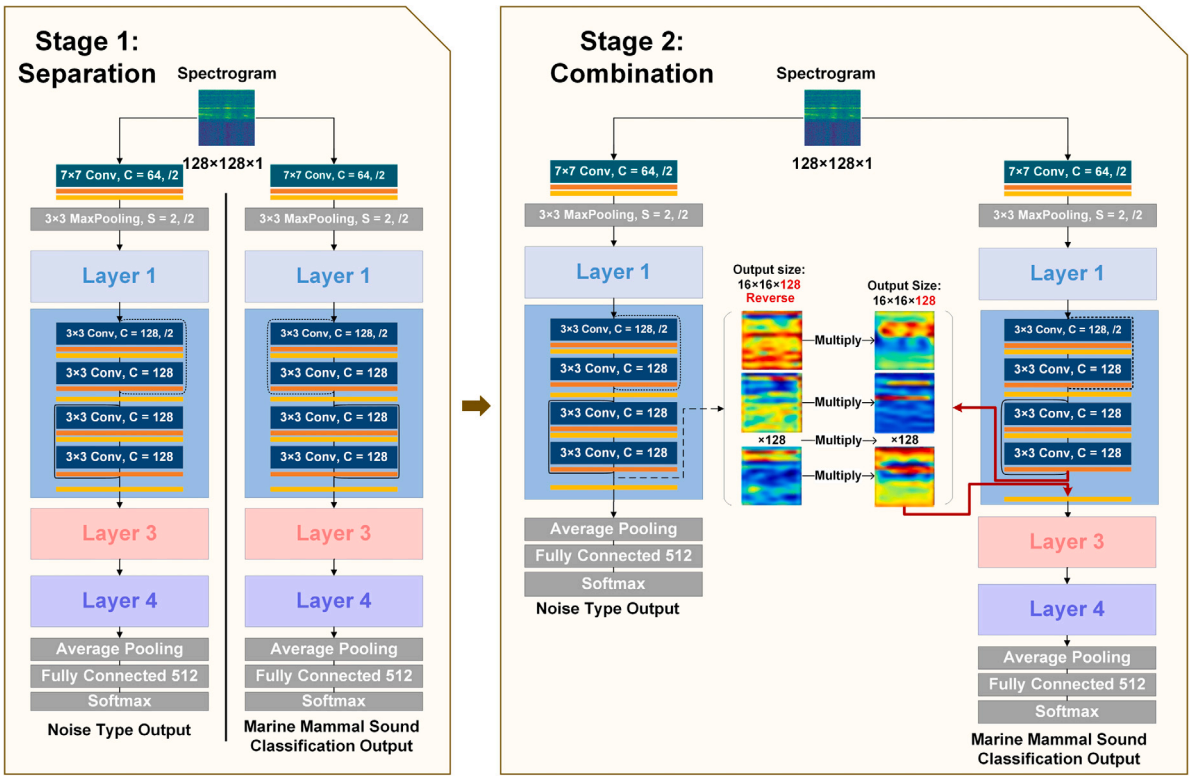


Fig. 3. Two-stage training process.

Where L represents the loss function. Even when the residual gradient $\frac{\partial L}{\partial y}$ approaches zero, the identity term guarantees non-zero gradient flow to earlier layers, effectively addressing the vanishing gradient problem. ResNet18 has demonstrated strong performance in image and speech recognition tasks due to its balance between depth, representational power, and computational efficiency. In this work, log-scaled spectrograms serve as input, enabling end-to-end training for marine mammal sound classification.

3.3. MT-MaskNet framework

To enhance noise robustness and improve data utilization, we propose MT-MaskNet, a MTL framework built on ResNet18 as shown in Fig. 3. Marine mammal sound classification is designated as the primary task, while noise-type classification serves as the auxiliary task. A lightweight gating mechanism, integrated at the intermediate feature level, dynamically suppresses noise-related activations while amplifying target acoustic cues.

We employ a two-stage training strategy to decouple sound and noise representations. In Stage 1, both branches learn discriminative features independently using task-specific labels, which prevents gradient conflicts and minimizes negative transfer. In Stage 2, the gating module is activated: features from Layer 2 of the sound branch $\mathbf{F}_{main}$ and noise branch $\mathbf{F}_{aux}$ are extracted. The auxiliary branch generates a suppression mask $\mathbf{G} \in \mathbb{R}^{C\times H\times W}$ The auxiliary branch generates a suppression mask $\mathbf{G}$ and $\mathbf{F}_{main}$ yields the gated features:

$$\mathbf{F}_{gated} = \mathbf{G} \odot \mathbf{F}_{main} \tag{3}$$

The gated features are then propagated through the subsequent layers of the primary branch to produce the final classification output. We apply the gating mechanism at Layer 2 ($128 \times 16 \times 16$ resolution), as this intermediate stage optimally balances fine-grained time–frequency details with abstract contextual patterns. This specific placement, as validated in Section 4.4, prevents the loss of critical signal cues that often occurs in deeper, more abstracted layers. The total loss function combines the cross-entropy losses from both tasks in a weighted manner:

$$L_{total} = \omega_{sound} \cdot L_{sound} + \omega_{noise} \cdot L_{noise} \tag{4}$$

Here, ω_{sound} and ω_{noise} are hyperparameters empirically tuned through ablation studies to optimize primary-task performance. During Stage 2, the auxiliary branch is truncated after Layer 2 to minimize computational overhead, while preserving its function in generating the gating signal. This architecture facilitates end-to-end training, obviating the requirement for independent denoising steps and allowing for adaptive noise suppression based on explicit noise-type supervision, thereby enhancing the model's ability to disentangle sound and interference in a unified manner.

4. Experiments and results

4.1. Datasets

The primary dataset comprised the 'Best Cut' of sounds from three dolphin species: White-Beaked Dolphin, Atlantic-Spotted Dolphin, and White-Sided Dolphin extracted from the Watkins Marine Mammal Sound Database. These three species were selected due to their relatively abundant high-quality samples and largely overlapping geographical distributions in the North Atlantic. Fig. 4 illustrates the recording locations and sample durations for each species, providing context for the geographical overlap in the North Atlantic. After preprocessing (down-sampling to 60.6 kHz and segmentation), all segments were pooled, randomly shuffled, and partitioned into training, validation, and test sets using an 8:1:1 ratio to form the raw dataset.

To reduce overfitting and improve generalization, a data augmentation strategy based on additive noise synthesis is adopted. Following Nanni et al. (2020), clean marine mammal sounds are overlaid with realistic ambient noise to simulate diverse underwater acoustic conditions. Noise samples are sourced from the SanctSound project, which provides long-term, high-quality ocean soundscape recordings

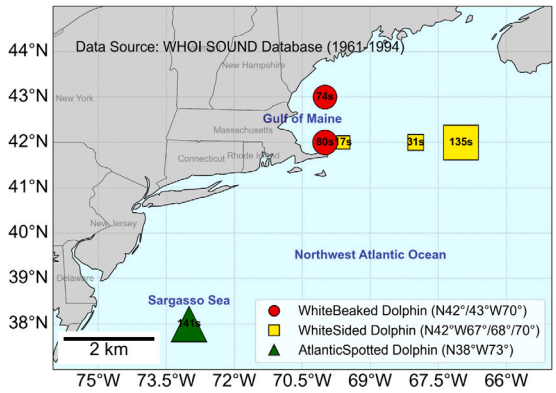


Fig. 4. Geographical distribution of recording locations and sample durations for selected dolphin species.

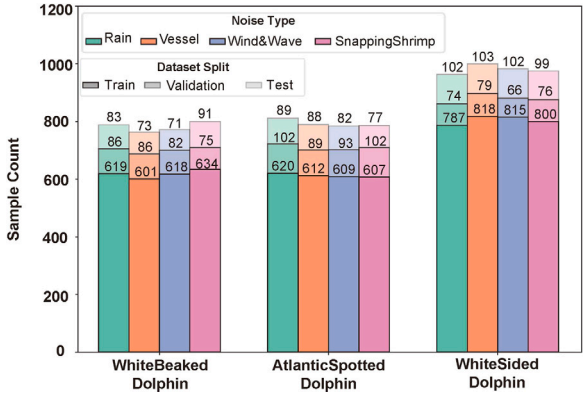


Fig. 5. Sample distribution by species and noise-type (stacked by split).

from eight marine protected areas worldwide. Four representative noise types are selected: rainfall, wind&wave, vessel, and snapping shrimp. All noise recordings are down-sampled to 60.6 kHz and segmented into 0.5 s clips to match the temporal resolution of the clean dataset. Each clean sound is then mixed with these noise types at five controlled SNR levels: 0 dB, 5 dB, 10 dB, 15 dB, and 20 dB. The resulting synthesized augmentation dataset achieves a roughly balanced category distribution, as shown in Fig. 5.

For the real-world PAM test set, we additionally prepared a collection of long-duration field recordings: one 50s clip of White-Sided Dolphin from Ocean Conservation Research, one 50s of Atlantic-Spotted Dolphin from the Watkins database (excluded from the 'Best Cut' training subset), and one 26s clip of White-Beaked Dolphin from North Sailing. All real PAM recordings were down-sampled to 60.6 kHz to ensure consistency with the training data, and then segmented into 0.5 s non-overlapping clips using the same protocol as the training set.

4.2. Implementation details

To evaluate the effectiveness of the proposed MT-MaskNet for marine mammal sound classification, we conducted a series of comparative experiments. All audio samples were converted into log-scaled spectrograms with a fixed resolution of 128 × 128 pixels using STFT (Hann window, length N = 512, hop size R = 256, 50% overlap). These spectrograms served as input to the following classification models: AlexNet (Krizhevsky et al., 2012), VGG16 (Simonyan and Zisserman, 2015), GoogLeNet (Szegedy et al., 2014), ResNet18 (He et al., 2015),ViT (Dosovitskiy et al., 2021) and the proposed MT-MaskNet.

Experiments were implemented in Python 3.12 using the TensorFlow framework. Spectrograms were normalized via Z-score transformation, with statistics computed solely from the training set to prevent

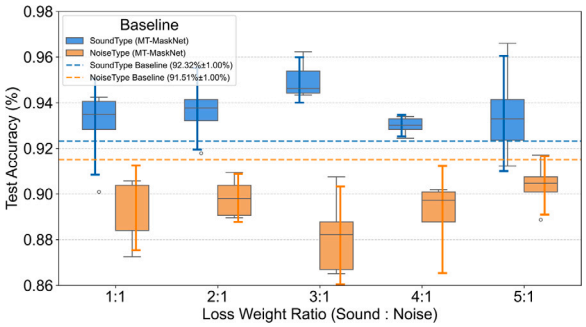


Fig. 6. Test accuracy distributions for soundtype and noisetype tasks with 95% confidence intervals across loss weight ratios (10 random seeds per ratio).

data leakage. Training was performed for up to 100 epochs with a batch size of 128, using the Adam optimizer and an initial learning rate of 1×10^{-3}.

4.3. Experimental results

Fig. 6 illustrates the test accuracy distributions for the SoundType (primary task) and NoiseType (auxiliary task) under varying loss weight ratios, along with their corresponding 95% confidence intervals, to demonstrate the optimal weighting that maximizes positive transfer while highlighting asymmetric task performance. MT-MaskNet consistently surpasses the single-task ResNet18 baseline across all loss-weight configurations. Specifically, the 3:1 ratio yields the most pronounced improvement, where the primary task accuracy reaches its peak while maintaining a narrow confidence interval, indicating a robust positive transfer effect. In contrast, the NoiseType task exhibits substantially lower performance compared to its single-task baseline in most configurations, with the most notable degradation at the 3:1 ratio.

This asymmetric performance reveals that the gating mechanism functions primarily as a noise-suppression filter rather than a precise classifier. By prioritizing the identification of noise-induced interference over fine-grained noise categorization, the model successfully reallocates its representational capacity to benefit the primary species-recognition task. The drop in auxiliary task performance at the 3:1 ratio likely reflects a successful reallocation of optimization resources toward the main task, allowing the gating to provide coarse-grained noise interference detection and feature suppression. This enables positive transfer to SoundType by freeing up representational capacity, while inducing negative transfer to NoiseType due to gradient competition under unbalanced weighting. Such a trade-off is a well-documented characteristic of MTL when tasks share parameters and compete for limited representational resources.

To rigorously assess whether the observed performance improvement of the MT-MaskNet over the single-task ResNet18 is attributable to the introduced architectural modifications rather than random fluctuations, a statistical significance test is essential. As shown in Table 1, which presents the paired t-test results to quantify the statistical significance of performance differences, the proposed multi-task method achieved a mean test accuracy of 95.00% ± 0.80%, compared to 92.32% ± 1.10% for the single-task baseline. A paired t-test confirmed a improvement ($t(9) = 7.364$, $p = 0.0018$, two-tailed), with an average gain of 2.68% (95% CI: [1.53%, 3.83%]). These results demonstrate that the proposed approach yields a robust and significant performance advantage over the baseline under matched experimental conditions.

Fig. 7 compares the Precision, Recall, and F1-Score distributions across three marine mammal species evaluated over ten independent random seeds. The proposed method consistently outperforms the ResNet18 baseline across all metrics, with the most notable improvements observed in F1-Score, indicating a more balanced trade-off

Table 1
Paired t-test results comparing test accuracy of MT-MaskNet and ResNet18 across 10 random seeds.

Metric	Proposed method (Mean ± SD)	Baseline (Mean ± SD)	Mean difference	95% CI of difference	$t(9)$	p-value (two-tailed)
Test Accuracy (%)	95.00 ± 0.80	92.32 ± 1.10	+2.68	[1.53, 3.83]	7.364	0.0018

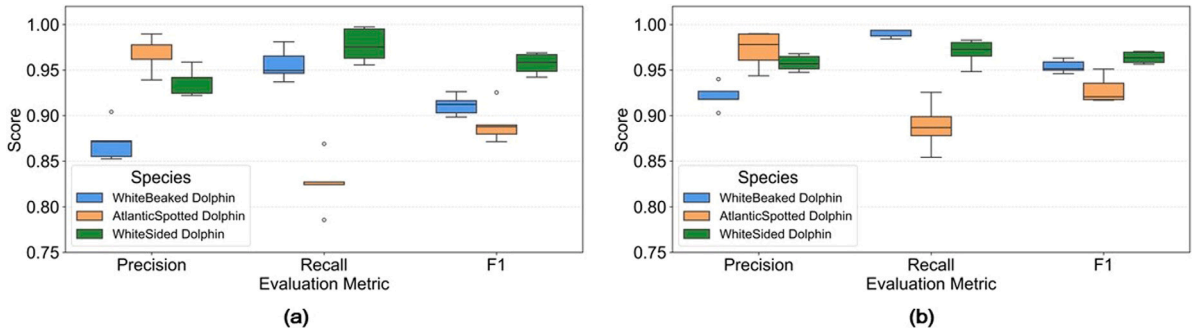


Fig. 7. Precision, recall, and F1-score of three dolphin species evaluated across ten independent random seeds. (a) shows the performance of the single-task baseline ResNet18. (b) presents the results of the MT-MaskNet under the optimal loss weight ratio of 3:1. Boxes represent the interquartile range, horizontal lines indicate medians, and whiskers extend to the minimum and maximum values (excluding outliers).

Table 2
Species-level performance comparison of single-task ResNet18 and proposed MT-MaskNet on real-world PAM data.

Species	Accuracy (%)		Precision (%)		Recall (%)		F1-Score (%)	
	ResNet18	MT-MaskNet	ResNet18	MT-MaskNet	ResNet18	MT-MaskNet	ResNet18	MT-MaskNet
White-Beaked Dolphin			48.54	85.71	100.00	97.50	65.36	91.23
Atlantic-Spotted Dolphin	44.00	73.75	00.00	59.63	00.00	81.25	00.00	68.78
White-Sided Dolphin			100.00	85.00	32.00	42.50	48.48	56.67

between precision and recall. For performance on White-Beaked Dolphin, the baseline exhibits high recall but relatively low precision, suggesting a moderate false positive rate. The proposed method substantially improves precision while maintaining strong recall, thereby effectively reducing misclassification errors. Performance on Atlantic-Spotted Dolphin, which shows the lowest and most variable recall in the baseline, benefits the most from the gating mechanism, with improved recall and sustained high precision, resulting in the largest F1-score gain and reduced false positives and false negatives. The ResNet18 already performs well in the classification of White-Sided Dolphin, yet the proposed method further enhances precision and stability, yielding higher F1-scores.

These results demonstrate that the gating mechanism enhances feature representation by suppressing noise-related interference, leading to lower false positive rates (higher precision) across all species, particularly alleviating the misclassification challenges observed for Atlantic-Spotted Dolphin, while preserving or improving recall, ultimately achieving more robust and balanced classification performance in the multi-task setting.

Table 2 presents a species-level performance breakdown on real-world PAM data, highlighting the model's ability to address species-specific challenges in authentic underwater acoustic environments and demonstrating substantial improvements in robustness across diverse noise conditions. Overall, the baseline ResNet18 achieves a mean accuracy of 44.00%, which MT-MaskNet substantially improves to 73.75%.

The baseline exhibits highly unstable and species-dependent performance. For White-Beaked Dolphin, it achieves a perfect recall of 100%, meaning that the model correctly identifies every true instance of this species' sounds, but is hampered by low precision (48.54%), resulting in a high false positive rate — many non-target sounds are incorrectly classified as White-Beaked Dolphin — and an F1-score of 65.36%. Atlantic-Spotted Dolphin shows complete failure across all metrics (00.00), indicating severe masking of its acoustic features by ambient noise. In contrast, White-Sided Dolphin displays perfect precision of 100%, meaning that every prediction made for this species is correct, but extremely low recall of 32.00%, reflecting excessive conservatism that leads to frequent missed detections and a modest F1-score of 48.48%.

In comparison, MT-MaskNet delivers substantial and consistent improvements across all three species. For White-Beaked Dolphin, precision increases markedly to 85.71%, with recall remaining high at 97.50%, yielding an improved F1-score of 91.23% and a clear reduction in false positives. The most pronounced recovery is observed for Atlantic-Spotted Dolphin, where all metrics rise from zero to substantial values (precision 59.63%, recall 81.25%, F1-score 68.78%), demonstrating the gating mechanism's effectiveness in recovering discriminative features under noise interference. For White-Sided Dolphin, MT-MaskNet achieves a more balanced trade-off by elevating recall from 32% to 42.5% while preserving relatively high precision 85%, resulting in an enhanced F1-score of 56.67%.

These species-level results underscore that the proposed gating-based multi-task framework not only outperforms the single-task baseline in authentic PAM settings but also mitigates pronounced inter-species disparities through adaptive noise suppression and robust feature enhancement, thereby promoting lower false positive rates, reduced missed detections, and more equitable classification performance across diverse marine mammal sounds.

4.4. Ablation studies on gating mechanisms

Table 3 systematically evaluates the impact of the gating mechanism's placement and design variants within the ResNet18 backbone through a series of ablation studies, revealing performance on both the synthesized test set (under optimal loss weights) and real-world PAM data. The Layer2+Gating configuration attained peak performance in both synthesized and real-world environments. This superiority confirms that early-layer intervention is critical for noise robustness, as it intercepts interference before it propagates into the more abstracted semantic spaces of deeper layers. Specifically, early-layer gating (e.g., Layer1 and Layer2) substantially outperforms later

Table 3
Ablation on gating layer placement and mechanism type.

	Test accuracy (best loss weights configuration)	Accuracy on real PAM data (%)
Layer1+Gating	93.81% ± 0.76% (3:1)	71.33%
Layer2+Gating (Proposed)	95.00% ± 0.72% (3:1)	73.75%
Layer2+Adding	93.74% ± 1.11% (5:1)	62.50%
Layer3+Gating	92.85% ± 0.83% (1:1)	56.00%
Layer4+Gating	93.11% ± 1.00% (2:1)	44.17%

Table 4
Computational efficiency and performance comparison of baseline models and proposed MT-MaskNet on real PAM data.

	Models	Real-time factor (RTF)	Total parameters	Accuracy on real PAM data
CNN-based	AlexNet	0.0034	114.00 MB	0.3417
	VGG11	0.0090	22.72 MB	0.4375
	GoogleNet	0.0056	206.58 MB	0.4375
	ResNet18	0.0066	43.57 MB	0.4400
Transformer-based	ViT-Tiny	0.0097	34.16 MB	0.4958
	ViT-Small	0.0202	216.97 MB	0.5000
Proposed	MT-MaskNet	0.0086	46.42 MB	0.7375

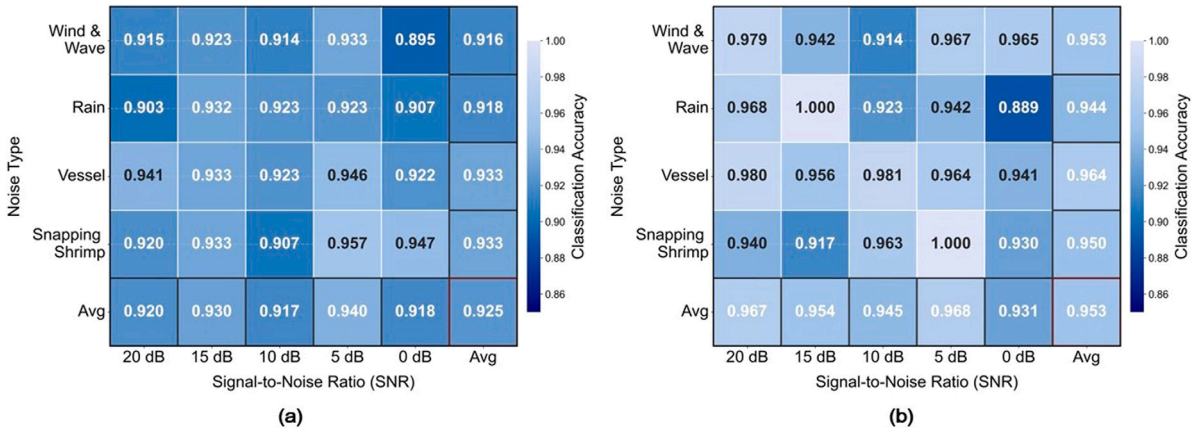


Fig. 8. Heatmap comparison of classification accuracy across different noise types and SNR levels. (a) single-task ResNet18 baseline, (b) proposed MT-MaskNet. Color gradients range from deep blue (indicating low accuracy) to light blue/white (high accuracy, up to 1.000).

placements (e.g., Layer3 and Layer4), likely because early features remain closer to the raw time–frequency representations, enabling the coarse-grained auxiliary noise cues to guide more effective initial suppression of broad interference patterns, whereas later features are highly abstracted and semantically rich, making gating more prone to introducing unnecessary perturbations or over-suppression and leading to degraded accuracy (with Layer4 yielding the lowest performance). Furthermore, substituting the multiplicative gate with simple addition (Layer2+Adding) led to a performance decline. This drop underscores that dynamic, element-wise suppression is indispensable for the selective filtering of noise, whereas additive fusion indiscriminately introduces interfering features. The overall accuracy decline on real PAM data reflects inherent generalization challenges from synthesized to authentic environments, yet the proposed configuration maintains the top performance, indicating its promising potential for complex marine soundscapes.

4.5. Comparison with baseline models

As shown in Table 4, the proposed MT-MaskNet achieves the highest accuracy on real-world PAM recordings, substantially outperforming all baseline models. Notably, while Transformer-based models such as ViT-Tiny and ViT-Small attain higher single-task accuracies than the ResNet18 baseline, they exhibit clear limitations in practical PAM deployments. ViT-Tiny incurs a higher real-time factor (RTF) compared to ResNet18, while ViT-Small suffers from higher computational overhead, rendering both less suitable for energy-constrained, long-term underwater monitoring systems.

Integrating the MT-MaskNet framework with ResNet18 yields a synergy, driving a remarkable performance leap from 44.00% to 73.75% on real-world recordings. This 29.75% absolute gain highlights the framework's robust generalization and effectively bridges the gap between synthetic training and field applications. Furthermore, this substantial improvement comes with minimal computational cost, as the model maintains a low RTF of 0.0086 and only a marginal parameter increase (+6.6%, 46.42 MB). In contrast, Transformer models offer diminishing returns in this short-clip setting, despite their global modeling strengths in longer-sequence scenarios (typically 2–5 s in prior studies Cotillard et al., 2024; Li et al., 2024; Makropoulos et al., 2025; Zhang et al., 2025). These results underscore that for real-time, resource-limited PAM, specialized noise-robustness mechanisms and domain-specific architectural alignment outweigh the benefits of global attention alone.

4.6. Robustness to noise levels and types

To further evaluate the robustness of MT-MaskNet under diverse noise conditions, we compared its classification performance against that of ResNet18 across multiple noise scenarios. Fig. 8 provides a comparative heatmap visualization of classification accuracy across various noise-types and SNR levels. Overall, MT-MaskNet achieves a global improvement in average accuracy, elevating it from 0.925 in

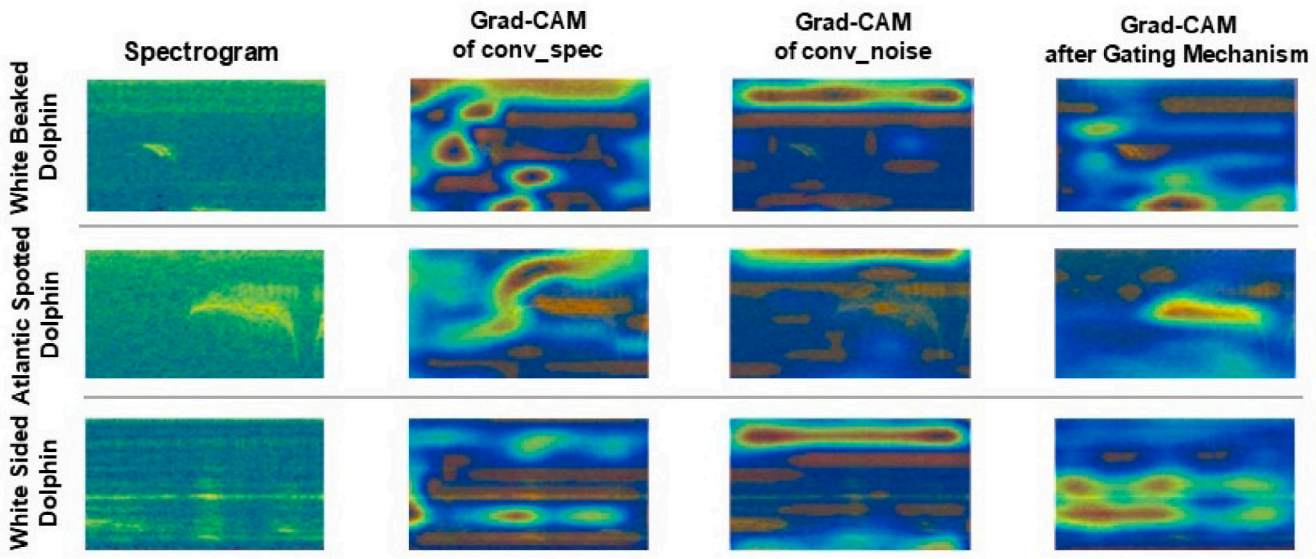


Fig. 9. Grad-CAM visualizations of feature activations in MT-MaskNet before and after the gating mechanism.

the baseline to 0.953—a relative gain of approximately 3.0%. This underscores the efficacy of the MTL framework integrated with the gating mechanism, which adaptively suppresses noise activations and preserves salient sound features, leading to more stable and superior outcomes in the majority of tested scenarios.

However, limitations are evident in isolated cases, where MT-MaskNet exhibits marginal degradation compared to the baseline, notably under snapping shrimp noise at 15 dB SNR (0.917 vs. 0.933) and 0 dB SNR (0.930 vs. 0.947), as well as rain noise at 0 dB SNR (0.889 vs. 0.907). These instances suggest potential over-suppression or suboptimal masking in specific noise profiles, highlighting the need for further refinement in handling pulse-like or broadband interferences to ensure comprehensive robustness.

4.7. Feature visualization and interpretability

Fig. 9 illustrates the Grad-CAM visualizations of feature activations within the MT-MaskNet architecture for three representative test samples, to provide interpretative evidence of how the gating mechanism adaptively suppresses noise-induced features and enhances target sound patterns. From left to right, the columns depict: (1) the original spectrogram, capturing the time–frequency representation of the input signal; (2) Grad-CAM heatmaps of the convolutional activations from the sound classification branch (conv_sound) prior to gating, highlighting regions where the model focuses on sound-specific patterns; (3) Grad-CAM heatmaps from the noise classification branch (conv_noise), emphasizing areas dominated by ambient interference; and (4) the post-gating activations after element-wise multiplication, demonstrating the adaptive suppression of noise-induced features. Grad-CAM visualizations demonstrate that pre-gating activations in the sound branch concentrate on salient vocal contours. Simultaneously, the noise branch identifies broad interference zones, allowing the subsequent gating operation to suppress these regions and sharpen the model's focus on target acoustic cues. After gating, the heatmaps exhibit a more refined focus. Spurious activations are markedly reduced, and contrast on target features is enhanced. These changes demonstrate the gating mechanism's effectiveness in strengthening discriminative representations for robust classification in noisy underwater environments.

5. Discussion

5.1. Mechanistic role of the gating mechanism

The empirical evidence from Grad-CAM visualizations and asymmetric task performance further elucidates the mechanistic role of the gating mechanism. Specifically, under the optimal loss weight ratio, the primary sound classification task exhibits substantial gains while the auxiliary noise classification task shows degraded performance. This asymmetry indicates that the gating module, applied at Layer 2, primarily functions as a coarse-grained noise indicator rather than enabling precise noise categorization. It effectively suppresses broadband interference activations in early layers while preserving salient acoustic patterns of target sounds. Such behavior aligns with established MTL principles, whereby auxiliary tasks can serve as regularizers that facilitate beneficial feature sharing and resource reallocation, even at the cost of their own accuracy (Crawshaw, 2020). Such coarse guidance is especially pertinent to bio-acoustic applications, as the auxiliary branch can generate effective suppression signals for the primary task without requiring high classification precision.

5.2. Comparison with modern architectures and masking techniques

The proposed learnable feature-level gating offers distinct advantages over conventional noise-masking strategies. Unlike preprocessing-based IBM (Olatinwo and Seto, 2024) or attention-based masks (Razig et al., 2025), which often risk early-stage information loss and incur extra computational overhead, MT-MaskNet integrates suppression directly within the latent representations in an end-to-end manner.

Furthermore, while Transformer-based architectures excel at global modeling for long sequences (typically 2–5 s in prior studies Cotillard et al., 2024; Li et al., 2024; Makropoulos et al., 2025; Zhang et al., 2025), the selection of a CNN backbone for MT-MaskNet remains more appropriate given the temporal scale of our dataset. For the 0.5 s clips, ResNet18's localized receptive fields provide a more efficient inductive bias than the quadratic self-attention mechanism of Transformers. This synergy ensures high classification accuracy while maintaining a low RTF of 0.0086, thereby optimizing the model for energy-constrained hardware deployments.

Critically, MT-MaskNet provides clearer physical interpretability than existing MTL frameworks used in underwater target recognition (Huang et al., 2025; Li et al., 2023). Whereas prior works often depend on implicit feature sharing across task branches, our gating mechanism generates an explicit, visualizable suppression signal. This transparency, substantiated by Grad-CAM heatmaps, reveals precisely how noise-related activations are attenuated to emphasize salient acoustic cues.

5.3. The sim-to-real gap and generalization challenges

Although our results underscore the model's robustness under controlled noise, the performance disparity on real-world datasets reveals a pervasive Sim-to-Real gap inherent in PAM. The observed decline — particularly the sharp reduction in White-Sided Dolphin recall — echoes the generalization challenges documented by Hamard et al. (2024) regarding out-of-distribution click sequences. This divergence suggests that linear additive synthesis acts as an idealized proxy that underrepresents the stochastic complexity of the marine channel. Beyond simple superposition, authentic field recordings encapsulate environment-induced modulations, such as frequency-selective attenuation and multipath distortions, alongside systematic sensor biases that are absent in laboratory-mixed datasets. To mitigate these challenges, future efforts should prioritize unsupervised domain adaptation (UDA) (Doig et al., 2025) and hybrid generative-synthetic strategies (Padovese et al., 2025) to ensure model reliability across diverse and evolving oceanic soundscapes.

5.4. Limitations and future scalability

The proposed framework also shows strong potential for extension to weakly supervised or unsupervised settings. Techniques such as contrastive learning or autoencoders in the auxiliary branch (Acs et al., 2026; Bermant et al., 2022) could learn latent noise representations from unlabeled data. This capability would greatly improve scalability in large-scale PAM deployments where comprehensive noise-type labels are sometimes unavailable. However, noisy or incomplete labels may still introduce suboptimal gating and gradient conflicts, underscoring the need for the uncertainty-aware training mechanisms.

Beyond label constraints, the taxonomic scope of this study was deliberately limited to three representative dolphin species to maintain rigorous experimental control. While this focus ensured statistical reliability, future work must evaluate the framework on larger, multi-species datasets to encompass a broader breadth of marine mammal sounds. Furthermore, exploring hybrid CNN–Transformer architectures may address the need for longer temporal contexts, while field validation on operational platforms — such as autonomous gliders and buoys — remains essential to bridge the remaining sim-to-real gap. The balance of accuracy, efficiency, and interpretability established here suggests that MT-MaskNet is well-suited for resource-constrained environments, providing a scalable foundation for advanced marine mammal conservation and ecosystem management.

6. Conclusion

This study introduces MT-MaskNet, an MTL framework that integrates noise-type classification as an auxiliary task with a mid-level gating mechanism to enable adaptive, feature-level suppression. By employing a two-stage training strategy, the model effectively disentangles sound and noise representations, thereby mitigating negative transfer while maximizing performance gains for the primary classification task. Evaluations on both synthesized and real-world recordings demonstrate that coarse-grained gating guidance significantly enhances robustness, particularly in challenging underwater environments where target signals are masked by non-stationary noise. This mechanistic approach not only obviates the need for independent denoising modules but also maintains high computational efficiency, evidenced by a low RTF suitable for resource-constrained deployments. While a performance gap persists between synthesized and authentic data due to complex oceanic propagation effects, the interpretability and efficiency of MT-MaskNet provide a scalable foundation for automated marine mammal monitoring. Future research will focus on bridging the remaining sim-to-real shift through UDA and expanding the framework's taxonomic scope across more diverse acoustic soundscapes.

CRediT authorship contribution statement

Xuchen Wang: Writing – original draft, Visualization, Validation, Methodology. **Yougan Chen:** Writing – review & editing. **Kunyun Du:** Writing – review & editing. **Yihao Zhao:** Writing – review & editing. **Yanhan Dong:** Writing – review & editing. **Shen'ao Tu:** Writing – review & editing. **Yi Tao:** Writing – review & editing. **Xiaomei Xu:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Yougan Chen reports financial support was provided by National Natural Science Foundation of China. Yougan Chen reports a relationship with National Natural Science Foundation of China that includes: funding grants.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62271423, in part by the Basic Research Program of Science and Technology of Shenzhen, China under Grant JCYJ20230807091406013, in part by the State Key Laboratory of Acoustics and Marine Information, Chinese Academy of Sciences under Grant SKLA202505, and in part by the Fujian Ocean Innovation Center under Grant 25FV0CZJ03.

Data availability

The data samples generated and analyzed, along with the complete source code for the MT-MaskNet model, are publicly available in the GitHub repository. The data and code can be accessed via the following permanent link: https://github.com/XC4Marine/MT-MaskNet.

References

Acs, R., Ibrahim, A., Zhuang, H., Chérubin, L.M., 2026. Contrastive learning for passive acoustic monitoring: A framework for sound source discovery and cross-site comparison in marine soundscapes. PLoS Comput. Biol. 22 (3), e1014005. http://dx.doi.org/10.1371/journal.pcbi.1014005.

Aslam, M., Zhang, L., Liu, X., Irfan, M., Xu, Y., Li, N., Zhang, P., Jiangbin, Z., Yaan, L., 2024. Underwater sound classification using learning based methods: a review. Expert Syst. Appl. 255, 124498. http://dx.doi.org/10.1016/j.eswa.2024.124498.

Atito, S., Awais, M., Wang, W., Plumbley, M.D., Kittler, J., 2024. ASiT: Local-global audio spectrogram vIsion transformer for event classification. IEEE/ACM Trans. Audio, Speech, Lang. Process. 32, 3684–3693. http://dx.doi.org/10.1109/TASLP.2024.3428908.

Bermant, P.C., Brickson, L., Titus, A.J., 2022. Bioacoustic event detection with self-supervised contrastive learning. http://dx.doi.org/10.1101/2022.10.12.511740, BioRxiv, 2022.10.12.511740.

Bermant, P.C., Bronstein, M.M., Wood, R.J., Gero, S., Gruber, D.F., 2019. Deep machine learning techniques for the detection and classification of sperm whale bioacoustics. Sci. Rep. 9 (1), 12588. http://dx.doi.org/10.1038/s41598-019-48909-4.

Cai, W., Zhu, J., Zhang, M., Yang, Y., 2022. A parallel classification model for marine mammal sounds based on multi-dimensional feature extraction and data augmentation. Sensors 22, 7443. http://dx.doi.org/10.3390/s22197443.

Cauchy, P., Heywood, K., Merchant, N., Risch, D., Queste, B., Testor, P., 2023. Gliders for passive acoustic monitoring of the oceanic environment. Front. Remote. Sens. 4, 1106533. http://dx.doi.org/10.3389/frsen.2023.1106533.

Cotillard, T., Sécheresse, X., Aubin, J., Mikus, M.A., Vergara, V., Gambs, S., Michaud, R., Martins, C.C.A., Turgeon, S., Chion, C., Roca, I., 2024. Automatic detection and classification of beluga whale calls in the St. Lawrence estuary. J. Acoust. Soc. Am. 156 (6), 3723–3740. http://dx.doi.org/10.1121/10.0030472.

Crawshaw, M., 2020. Multi-task learning with deep neural networks: a survey. URL: https://arxiv.org/abs/2009.09796, arXiv:2009.09796.

Deng, S., Hong, F., 2025. Advancing underwater acoustic target recognition in low-SNR environments with UATR-DIFF-transformer. Ocean Eng. 341, 122668. http://dx.doi.org/10.1016/j.oceaneng.2025.122668.

Doig, H., Pizarro, O., Williams, S., 2025. Training marine species object detectors with synthetic images and unsupervised domain adaptation. Front. Mar. Sci. 12, http://dx.doi.org/10.3389/fmars.2025.1581778.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An image is worth 16x16 words: Transformers for image recognition at scale. URL: https://arxiv.org/abs/2010.11929, arXiv:2010.11929.

Fei, N., Gao, Y., Lu, Z., Xiang, T., 2021. Z-score normalization, hubness, and few-shot learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. ICCV, http://dx.doi.org/10.1109/ICCV48922.2021.00021.

Gu, A., Gulcehre, C., Paine, T., Hoffman, M., Pascanu, R., 2020. Improving the gating mechanism of recurrent neural networks. URL: https://arxiv.org/abs/1910.09890, arXiv:1910.09890.

Hamard, Q., Pham, M.T., Cazau, D., Heerah, K., 2024. A deep learning model for detecting and classifying multiple marine mammal species from passive acoustic data. Ecol. Informatics 84, 102906. http://dx.doi.org/10.1016/j.ecoinf.2024.102906.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep residual learning for image recognition. URL: https://arxiv.org/abs/1512.03385, arXiv:1512.03385.

Huang, W., Sun, S., Lu, J., Xu, Z., Xiu, Z., Zhang, H., 2025. A multi-task learning balanced attention convolutional neural network model for few-shot underwater acoustic target recognition. URL: https://arxiv.org/abs/2504.13102, arXiv:2504.13102.

Krizhevsky, A., Sutskever, I., Hinton, G., 2012. ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems. Curran Associates, Inc..

Lei, Z., Lei, X., Wang, N., Zhang, Q., 2022. Present status and challenges of underwater acoustic target recognition technology: a review. Front. Phys. 10, 1044890. http://dx.doi.org/10.3389/fphy.2022.1044890.

Li, D., 2023. Data augmentation method for underwater acoustic target recognition based on underwater acoustic channel modeling and transfer learning. Appl. Acoust. 208, 109344. http://dx.doi.org/10.1016/j.apacoust.2023.109344.

Li, D., Liao, J., Jiang, H., Jiang, K., Chen, M., Zhou, B., Pu, H., Li, J., 2024. A classification method of marine mammal calls based on two-channel fusion network. Appl. Intell. 54, 3017–3039. http://dx.doi.org/10.1007/s10489-023-05138-7.

Li, D., Liu, F., Shen, T., Chen, L., Zhao, D., 2023. A robust feature extraction method for underwater acoustic target recognition based on multi-task learning. Electronics 12, 1708. http://dx.doi.org/10.3390/electronics12071708.

Lü, Z., Shi, Y., Lü, L., Han, D., Wang, Z., Yu, F., 2024. Dual-feature fusion learning: an acoustic signal recognition method for marine mammals. Remote. Sens. 16, 3823. http://dx.doi.org/10.3390/rs16203823.

Makropoulos, D.N., Filntisis, P.P., Prospathopoulos, A., Kassis, D., Tsiami, A., Maragos, P., 2025. Improving classification of marine mammal vocalizations using vision transformers and phase-related features. In: 2025 25th International Conference on Digital Signal Processing. DSP, pp. 1–5. http://dx.doi.org/10.1109/DSP65409.2025.11075076.

Nanni, L., Maguolo, G., Paci, M., 2020. Data augmentation approaches for improving animal audio classification. Ecol. Informatics 57, 101084. http://dx.doi.org/10.1016/j.ecoinf.2020.101084.

Olatinwo, D., Seto, M., 2024. Detection of marine mammal vocalizations in low SNR environments with ideal binary mask. In: OCEANS 2024 - Halifax. pp. 1–6. http://dx.doi.org/10.1109/OCEANS55160.2024.10754093.

Olatinwo, D., Seto, M., 2025. Low-SNR northern right whale upcall detection and classification using passive acoustic monitoring to reduce adverse human-whale interactions. Mach. Learn. Knowl. Extr. 7 (4), 154. http://dx.doi.org/10.3390/make7040154.

Padovese, B., Frazao, F., Dowd, M., Joy, R., 2025. Advancing marine bioacoustics with deep generative models: a hybrid augmentation strategy for southern resident killer whale detection. URL: https://arxiv.org/abs/2511.21872, arXiv:2511.21872.

Palmero, S., Guidi, C., Kulikovskiy, V., Sanguineti, M., Manghi, M., Sommer, M., Pesce, G., 2023. Towards automatic detection and classification of orca (orcinus orca) calls using cross-correlation methods. Mar. Mam. Sci. 39 (2), 576–593. http://dx.doi.org/10.1111/mms.12990.

Peng, L., Lu, Z., Lei, T., Jiang, P., 2024. Dual-structure elements morphological filtering and local Z-score normalization for infrared small target detection against heavy clouds. Remote. Sens. 16, 2343. http://dx.doi.org/10.3390/rs16132343.

Razig, A., Soulaymani, Y., Benabbou, L., Cauchy, P., 2025. Multi-representation attention framework for underwater bioacoustic denoising and recognition. URL: https://arxiv.org/abs/2510.26838, arXiv:2510.26838.

Schäfer-Zimmermann, J.C., Demartsev, V., Averly, B., Dhanjal-Adams, K.L., Duteil, M., Gall, G., Faiß, M., Johnson-Ulrich, L., Stowell, D., Manser, M.B., Roch, M.A., Strandburg-Peshkin, A., 2026. Animal2vec and MeerKAT: A self-supervised transformer for rare-event raw audio input and a large-scale reference dataset for bioacoustics. Methods Ecol. Evol. 17 (3), 875–888. http://dx.doi.org/10.1111/2041-210x.70218.

Shiu, Y., Palmer, K., Roch, M., Fleishman, E., Liu, X., Nosal, E.M., Helble, T., Cholewiak, D., Gillespie, D., Klinck, H., 2020. Deep neural networks for automated detection of marine mammal species. Sci. Rep. 10, 607. http://dx.doi.org/10.1038/s41598-020-57367-y.

Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. URL: https://arxiv.org/abs/1409.1556, arXiv:1409.1556, Also published in ICLR 2015.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2014. Going deeper with convolutions. URL: https://arxiv.org/abs/1409.4842, arXiv:1409.4842, Also published in CVPR 2015.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30, http://dx.doi.org/10.48550/arXiv.1706.03762.

Vishnu, H., Soorya, V., Chitre, M., Too, Y., Koay, T., Ho, A., 2024. Machine-learning based detection of marine mammal vocalizations in snapping-shrimp dominated ambient noise. Mar. Environ. Res. 199, 106571. http://dx.doi.org/10.1016/j.marenvres.2024.106571.

White, E., White, P., Bull, J., Risch, D., Beck, S., Edwards, E., 2022. More than a whistle: automated detection of marine sound sources with a convolutional neural network. Front. Mar. Sci. 9, http://dx.doi.org/10.3389/fmars.2022.879145.

Zhang, Z., 2015. The singular value decomposition, applications and beyond. URL: https://arxiv.org/abs/1510.08532, arXiv:1510.08532.

Zhang, X., Liu, X., Alksne, M.N., Roch, M.A., 2025. Automating time × frequency annotations of delphinid whistles by adapting a foundational transformer neural network. Sci. Rep. 15 (1), 37809. http://dx.doi.org/10.1038/s41598-025-21642-x.