Bioacoustic Multi-Step Attention: Underwater Ecosystem Monitoring in Climate Change Context

Amine Razig[1,5], Youssef Soulaymani[2], Loubna Benabbou[3], Pierre Cauchy[4]
[1] MILA – Quebec AI Institute
[2] Université de Montréal
[3] Université du Québec à Rimouski
[4] ISMER
[5] Institut Polytechnique de Paris
amine.razig@polytechnique.edu

Abstract

Automated monitoring of marine mammals in the St. Lawrence Estuary faces extreme challenges: calls span low-frequency moans to ultrasonic clicks, often overlap, and are embedded in variable anthropogenic and environmental noise. We introduce a multi-modal, attention-guided framework that *first* segments spectrograms to generate soft masks of biologically relevant energy and *then* fuses these masks with the raw inputs for multi-band, denoised classification. Image and mask embeddings are integrated via mid-level fusion, enabling the model to focus on salient spectrogram regions while preserving global context. Using real-world recordings from the Saguenay–St. Lawrence Marine Park Research Station in Canada, we demonstrate that segmentation-driven attention and mid-level fusion improve signal discrimination, reduce false positive detections, and produce reliable representations for operational marine mammal monitoring across diverse environmental conditions and signal-to-noise ratios. By integrating attention-guided denoising with biodiversity-oriented evaluation metrics, our framework transforms raw hydrophone data streams into robust, operationally actionable presence signals, thereby supporting marine biodiversity conservation and climate-adaptation monitoring initiatives.

1 Introduction

The St. Lawrence Estuary is an acoustic habitat where protected marine mammal species must maintain essential biological functions, communication, navigation, and foraging, in the presence of increasing anthropogenic noise. Ship noise can mask calls and echolocation, disrupt essential behavioral sequences, and induce physiological stress[20] with ecosystem-level consequences when behaviors change over space and time. This acoustic degradation, exacerbated by the effects of climate change on marine soundscapes and species distributions, creates time-critical monitoring challenges that require robust automated detection systems capable of real-time assessment of species presence, behavioral state changes, and climate-driven population dynamics to inform adaptive conservation interventions. [22, 23]

These impacts have motivated concrete mitigation and policy efforts (e.g., quieter ship design, operational routing, and speed management) and targeted recovery planning for St. Lawrence species such as beluga. Our focus in this work is to turn raw hydrophone data into reliable presence signals that support biodiversity protection, monitoring, and adaptation actions in this sensitive region. **Our contributions:** First, we propose an end-to-end multi-step framework that segments spectrograms to produce pseudo attention masks and fuses mask and spectrogram embeddings to guide denoising and enhance biologically relevant signal recognition. Then we evaluate real-world recordings collected by

Tackling Climate Change with Machine Learning Workshop at NeurIPS 2025.

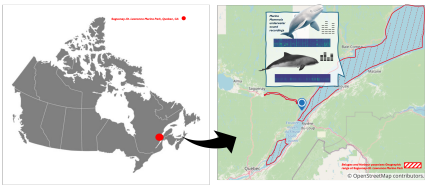


Figure 1: Saguenay–St. Lawrence Marine Park (SSLMP) representation.

the Saguenay–St. Lawrence Marine Park Research Station, emphasizing cross-season robustness and per-class precision, with control for empty signals. Finally, we demonstrate that segmentation-driven attention and mid-level fusion improve precision recall, stabilize detection thresholds, and produce robust field-ready representations for underwater bioacoustic monitoring.

2 Dataset description and problem setup

Dataset description We used an exclusive subset of the Saguenay - St. Lawrence Marine Park (SSLMP) monitoring dataset [7], a long-term multimodal collection designed to study the impact of maritime traffic on endangered marine mammals. Data come from two complementary sources: bottom-moored hydrophones (passive acoustic monitoring, PAM) that provide $\sim$ 1,500 hours of continuous recordings and shore-based surveys (LBS) that provide $\sim$ 500 hours of visual observations over four years. These data streams are synchronized, producing species-level annotations in [7] for belugas (*Delphinapterus leucas*) and harbour porpoises (*Phocoena phocoena*). Our subset consists of $\sim$10,000 five-minute segments manually annotated [7] with species presence and sound types (beluga whistles and clicks, 10–100 kHz; porpoise narrowband clicks, 50–150 kHz). The recordings also capture vessel noise and other natural and anthropogenic sounds spanning 10 Hz–150 kHz. The dataset is challenging due to environmental noise, overlapping calls, and domain shifts across seasons, sites, and sensors, making it a unique benchmark for machine learning in underwater bioacoustics.

Problem setup We work with a dataset of raw marine acoustic recordings containing vocalizations from multiple species. Our goal is to automatically recognize marine mammal vocalizations in noisy recordings, addressing challenges such as variable signal-to-noise ratios, overlapping calls, and environmental noise. We explore both multi-label and multi-class classification, before introducing attention mask driven framework using spectrogram-based representations of the audio data.

Formulation Formally, let $x(t)$ denote a raw acoustic waveform. The signal is first transformed into a spectrogram via a time-frequency representation (STFT). A segmentation model $\mathcal{M}_{\text{seg}}$ predicts a pseudo-attention mask highlighting relevant spectro-temporal regions. Both the spectrogram and the mask are then encoded into embeddings, which are fused to guide denoising and enhance biologically relevant signals. Finally, a classifier $\mathcal{C}$ maps the fused representation to the probabilities of the target class. Formally, the pipeline is:

$$\hat{y} = \mathcal{C}\Big(\text{Fuse}\Big(\mathcal{E}_{\text{spec}}\big(\mathcal{T}(x(t))\big),\ \mathcal{E}_{\text{mask}}\big(\mathcal{M}_{\text{seg}}(\mathcal{T}(x(t)))\big)\Big)\Big), \quad \hat{y} \in \mathbb{R}^K \tag{1}$$

where $\mathcal{T}$ is the STFT, $\mathcal{E}_{\text{spec}}$ and $\mathcal{E}_{\text{mask}}$ are the embedding functions for the spectrogram and mask, respectively, and $\text{Fuse}(\cdot,\cdot)$ denotes the mid-level embedding fusion.

3 Mask-driven classification method

Classification task The marine mammal acoustic signals were first analyzed by supervised classification in spectrogram representations capturing species-specific signatures. Two paradigms were considered. multi-class classification: and multi-label classification. We evaluated convolutional, modern CNN, and transformer-based architectures using standard metrics, applying ImageNet-based transfer learning [14]. Multi-class classification proved more suitable for our dataset, while noise and artifacts still limit the detection of subtle spectro-temporal patterns (see Fig. 6 and Tab. 3), motivating the denoising framework introduced next.

3.1 Automatic acoustic denoising framework

These difficulties discussed above can be largely attributed to noise that distorts the essential fine-grained temporal and spectral structures. To overcome these challenges, we introduce an automatic acoustic denoising framework designed to preprocess raw audio recordings prior to classification. This framework integrates signal transformation [2], mask-based denoising [1], and classification into a unified pipeline, thus improving robustness by clarifying relevant acoustic patterns through "pseudo-attention" masks and attention mechanisms.

Framework description Raw audio signals are first converted into time–frequency representations using the STFT. This operation decomposes the signal into overlapping windows. The resulting spectrograms are then used as the primary visual input for the denoising and classification stages. We apply a denoising methodology inspired by few-shot learning and leveraging the capabilities of models such as DeepLabV3 [21]. A substantial training set is constructed to train a segmentation model that generates "pseudo-attention" masks over spectrograms. These masks are then leveraged in a multi-step fusion framework, where both the raw spectrogram and its corresponding mask embedding are jointly encoded. The fused representation guides the network to focus on informative regions, effectively denoising the signal and enhancing underwater bioacoustic recognition. This approach is inspired by previous work in the audio denoising domain, notably the study on bird sounds [1], which demonstrated the effectiveness of deep visual denoising techniques in improving classification performance.

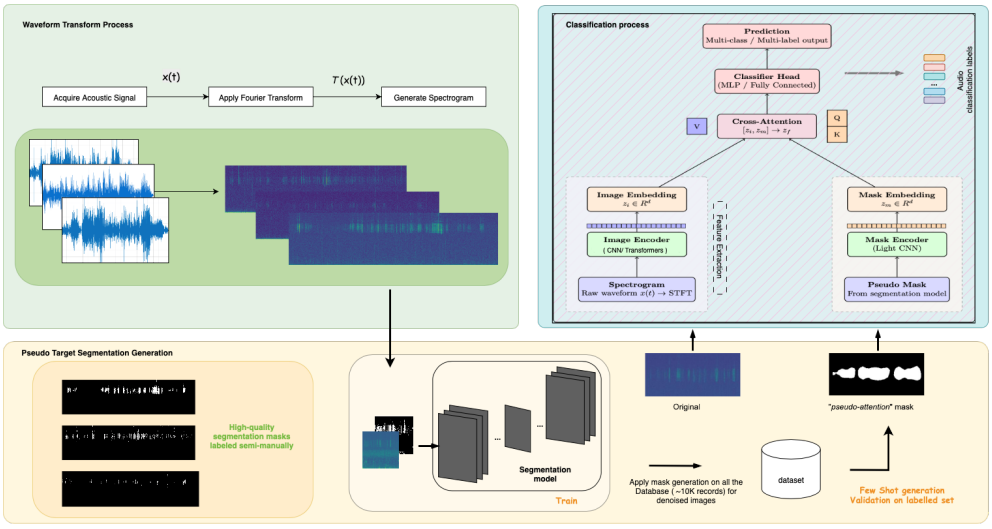


Figure 2: End-to-end framework for automatic denoising and classification from raw audio.

Audio transformation and Semi-automatic mask labelisation. The raw audio recordings are first converted to spectrogram representations using standard time-frequency analysis techniques. The spectrograms serve as the primary input for the subsequent denoising and classification stages. Once the spectrogram has been obtained, in order to efficiently annotate large collections, we adopt a semi-automatic labeling approach. First, an initial set of candidate regions is generated using signal processing techniques, such as edge detection and adaptive thresholding, to highlight potential patterns of interest. This allows us to identify and isolate prominent acoustic features. These preliminary masks are then presented to the annotator through an interactive interface, allowing manual refinement and correction, resulting in a high-quality training set (200 images) from which the denoising model can generalize mask predictions across the dataset.

Few-shot learning for denoising. Leveraging the high quality mammal sound pattern masks, we train a denoising model using a few-shot learning strategy to generalize from limited annotations. Architectures such as DeepLabV3 capture both fine-grained time–frequency structures and broader contextual patterns to distinguish signal from noise. In addition, we apply image horizontal flip augmentation to double the size of the training dataset. Once trained, the model predicts masks across the full dataset, enabling scalable denoising without exhaustive manual labeling.

Mask-guided model for classification. After training our segmentation model on spectrograms,

we obtain pseudo-attention masks that highlight regions most likely to contain relevant acoustic events. So, we threat it as an auxiliary representation [13]. Intuitively, the mask acts as a form of attention-based denoising: it emphasizes salient regions of the spectrogram while suppressing background noise and irrelevant structures. Concretely, we design a fusion framework with two parallel encoding branches: **Spectrogram encoder**, a ResNet50 or audio transformer backbone processes the raw spectrogram into a high-level representation. **Mask encoder**, a lightweight CNN encodes the corresponding segmentation mask into a compact embedding. Both embeddings are projected into a common latent space and then fused at an intermediate stage (mid-fusion). Fusion can be realized either by simple concatenation or through a cross-modal attention mechanism, where the spectrogram embedding serves as the query and the mask embedding provides keys and values. This enables the network to adaptively weigh spectro-temporal regions conditioned on the mask.Then, the fused representation is passed to a classification head, producing multi-class predictions. This design preserves a residual path from the spectrogram encoder to the classifier, ensuring that the system does not overly rely on potentially noisy masks while still exploiting their guidance signal. In doing so, we approximate the role of human attention in auditory scene analysis: focusing on the most informative patterns while filtering out distracting background components.

4 Results

4.1 Denoising process for marine mammals recognition

To evaluate the contribution of the proposed multi-step denoising framework, we compared it with standard image-only classification models trained on the same data set. Table 1 reports the accuracy and macro-F1 in ResNet50[11], ConvNeXt[10], ViT[12, 8], and our cross-attention fusion model using generated or high-quality (HQ) segmentation masks. In general, the results show that the multi-step approach substantially outperforms all baselines. Although ViT already provides strong performance among unimodal models (78. 8% accuracy), suggesting that attention mechanisms are better suited to model long-range temporal and spectral dependencies, the use of generated masks with cross-attention further improves the results to 83. 7%. The best performance is obtained with HQ masks (89.7% accuracy, 89.0% macro-F1), highlighting the benefit of leveraging accurate structural priors for denoising. This indicates that cross-attention enables the model to effectively exploit mask information to focus on relevant acoustic structures, and helps for the robustness of the classification.

Model	**Accuracy**	**F1 macro**
ResNet50	0.588	0.562
ConvNeXt	0.625	0.591
ViT	0.788	0.787
Multi-step (Gen. masks)	<u>0.837</u>	<u>0.816</u>
Multi-step (HQ masks)	**0.897**	**0.890**

Table 1: Comparison of baseline image-only models and the proposed multi-step approach with cross-attention using either generated or a **subset** with high-quality masks.

4.2 Ablation study of fusion methods

Fus. strategy	**High-Quality Masks**				**Generated Masks**			
	Train Loss	**Train Acc.**	**Val. Loss**	**Val. Acc.**	**Train Loss**	**Train Acc.**	**Val. Loss**	**Val. Acc.**
Concat	0.370	0.887	0.559	0.762	0.365	0.877	0.678	0.825
Gated	0.401	0.868	0.792	0.713	0.472	0.833	0.857	0.762
xAttn	0.253	0.912	0.406	**0.900**	0.427	0.843	0.695	**0.838**

Table 2: Comparison of mid-fusion strategies on the validation set using either high-quality (HQ) or generated (Gen.) masks. Cross-attention consistently achieves the best validation accuracy. (Training with RTX A100 GPU $\sim$ 15min per method)

We conducted an ablation study on the fusion strategy, comparing simple concatenation, gated residual fusion, and cross-attention; the results (Table 2) show that cross-attention achieves the best validation accuracy. These results suggest that, while simple and gated fusion capture some complementary information between the image and the mask but is more efficient with generated masks, introducing cross-attention enables more effective interaction between representations.

5 Conclusion

We presented a multi-step segmentation-based framework that improves the classification of marine mammal vocalizations using real-world data. While the use of STFT representations introduces resolution trade-offs and some information loss, the framework establishes a solid basis for robust and trustworthy ecological monitoring. Future work will address these limitations by exploring richer acoustic representations, improving attention mechanisms, and integrating predictive uncertainty. Overall, our results demonstrate that deep learning models can extract reliable presence signals that directly support species monitoring and conservation, illustrating how AI techniques can be effectively harnessed for scientific and climate-relevant ocean studies.

References

[1] Zhang, Y., Li, J. (2022). BirdSoundsDenoising: Deep Visual Audio Denoising for Bird Sounds. arXiv:2210.10196 [cs.SD].

[2] Xu, J., Xie, Y., Wang, W. (2024). Underwater Acoustic Target Recognition based on Smoothness-inducing Regularization and Spectrogram-based Data Augmentation. arXiv:2306.06945 [cs.SD].

[3] Jiang, Z., Soldati, A., Schamberg, I., Lameira, A. R., Moran, S. (2024). Automatic Sound Event Detection and Classification of Great Ape Calls using Neural Networks. arXiv:2301.02214 [eess.AS].

[4] Juodakis, J., Marsland, S. (2021). Wind-robust sound event detection and denoising for bioacoustics. arXiv:2110.05632 [stat.AP].

[5] Denton, T., Wisdom, S., Hershey, J. R. (2021). Improving Bird Classification with Unsupervised Sound Separation. arXiv:2110.03209 [eess.AS].

[6] Mishachandar, B., Vairamuthu, S. (2021). Diverse ocean noise classification using deep learning. Applied Acoustics, 181, 108141. doi:10.1016/j.apacoust.2021.108141.

[7] Bernier-Breton C. Écouter et observer les mammifères marins pour les étudier sans déranger: Approche combinée pour mieux comprendre l'utilisation de l'habitat par le béluga et le marsouin commun dans le parc marin du Saguenay–Saint-Laurent [thèse de maîtrise en océanographie]: Université du Québec à Rimouski; 2025.

[8] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H. (2020). Training data-efficient image transformers & distillation through attention. arXiv:2012.12877 [cs.CV].

[9] Sun, B., Luo, X. (2023). Underwater acoustic target recognition based on automatic feature and contrastive coding. IET Radar, Sonar & Navigation.

[10] Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., Xie, S. (2022). A ConvNet for the 2020s. arXiv:2201.03545 [cs].

[11] He, K., Zhang, X., Ren, S., Sun, J. (2015). Deep Residual Learning for Image Recognition. arXiv:1512.03385 [cs].

[12] 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. arXiv:2010.11929 [cs].

[13] Bayoudh, K., Knani, R., Hamdaoui, F., et al. (2022). A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. The Visual Computer, 38, 2939–2970. doi:10.1007/s00371-021-02166-7.

[14] Bengio, Y., Courville, A., Vincent, P. (2013). Representation Learning: A Review and New Perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(8), 1798-1828. doi:10.1109/TPAMI.2013.50.

[15] Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., He, Q. (2020). A Comprehensive Survey on Transfer Learning. arXiv:1911.02685 [cs].

[16] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 248-255. doi:10.1109/CVPR.2009.5206848.

[17] Gong, Y., Chung, Y. A., & Glass, J. (2021). AST: Audio spectrogram transformer. In Interspeech 2021 (pp. 571-575).

[18] Minyoung Huh, Pulkit Agrawal, & Alexei A. Efros. (2016) What makes ImageNet good for transfer learning? Berkeley Artificial Intelligence Research (BAIR) Laboratory

[19] Robin, O., Cauchy, P., Mercure-Boissonnault, P., Catineau, H., Mérindol, J., St-Onge, G., Gervaise, C., Gauthier-Marquis, J.-C., Kesour, K., Bazinet, M.-L., Lafrance, S. (2022) The MARS project: Identifying and reducing underwater noise from ships in the St. Lawrence Estuary. Canadian Acoustics, Vol. 50, No. 3.

[20] Erbe, C., Marley, S. A., Schoeman, R. P., Smith, J. N., Trigg, L. E., & Embling, C. B. (2019) The Effects of Ship Noise on Marine Mammals—A Review. Frontiers in Marine Science, 6(October).

[21] Chen, L.-C., Papandreou, G., Schroff, F., & Adam, H. (2017) Rethinking Atrous Convolution for Semantic Image Segmentation. arXiv preprint arXiv:1706.05587.

[22] D. P., Beger M., Boerder K., Boyce D. G., Cavanagh R. D., Cosandey-Godin A., et al. (2019). Integrating climate adaptation and biodiversity conservation in the global ocean. Sci. Adv. 5 (11).

[23] Laidre, K. L., Stern, H., Kovacs, K. M., Lowry, L., Moore, S. E., Regehr, E. V., ... Ugarte, F. (2015). Arctic marine mammal population status, sea ice habitat loss, and conservation recommendations for the 21st century: Arctic Marine Mammal Conservation. Conservation Biology, 29(3), 724–737.

6 Annexe

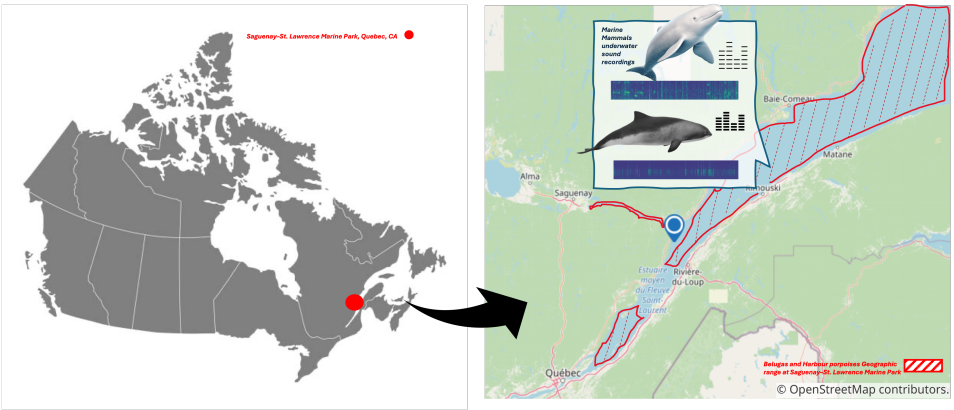


Figure 3: Saguenay–St. Lawrence Marine Park (SSLMP) representation.

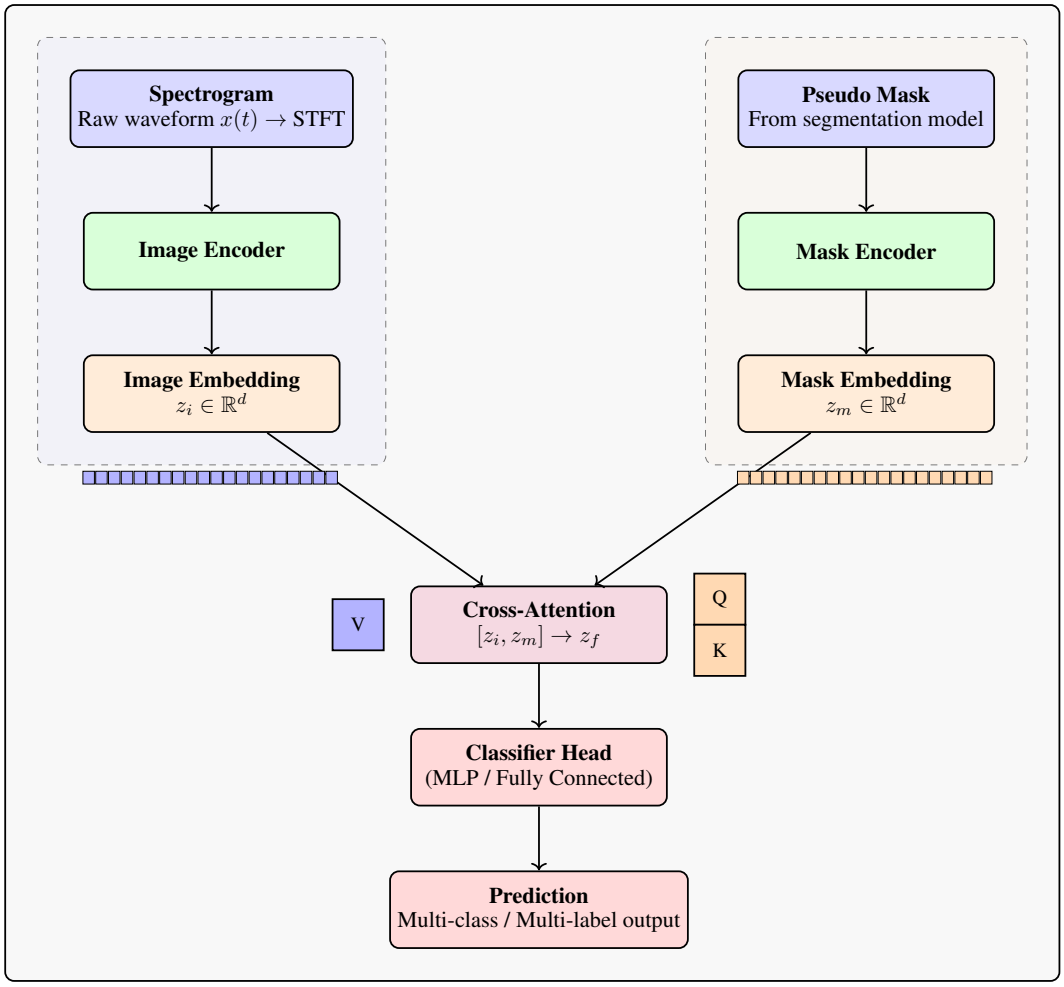


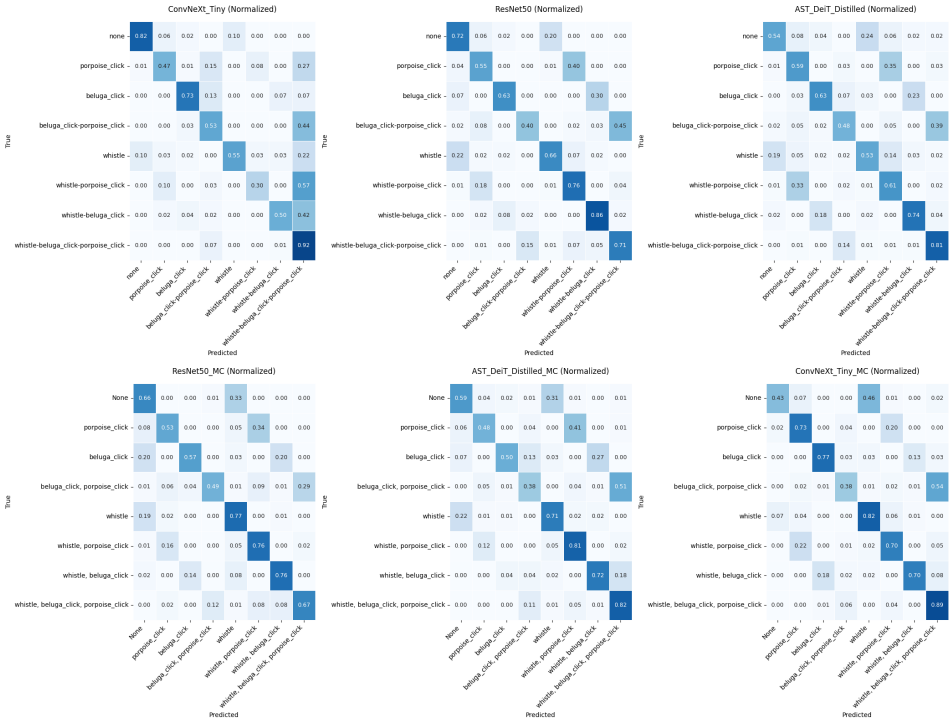
Figure 4: Architecture of the proposed model with two encoding branches and mid-fusion by cross-attention

Figure 5: Spectrogram (**left**), high-quality segmentation mask (**middle**), and generated pseudo-attention mask (**right**) for a recording of porpoise clicks.

Table 3: Performance comparison between multi-label and multi-class training approaches before multi-modal approach. *For multiclass (one label per sample): hamming loss is the average number of incorrect predictions per sample. For multilabel (multiple labels per sample): it is the average number of label errors per sample, divided by the number of labels. This metric is not comparable inter training method*

Metric	**ConvNeXt-Tiny**		**ResNet50**		**Deit-Distilled**	
	Multi-Label	Multi-Class	Multi-Label	Multi-Class	Multi-Label	Multi-Class
Hamming Loss	0.1693	0.3310	0.1206	0.3466	0.1427	0.3674
Perfect Accuracy	58.17%	**66.90%**	66.34%	65.34%	62.45%	63.26%
Whistle						
Precision	**0.806**	0.61	0.745	0.60	0.730	0.64
Recall	**0.891**	0.82	0.816	0.77	0.745	0.71
F1-Score	**0.847**	0.70	0.779	0.68	0.737	0.67
Beluga Click						
Precision	0.672	0.68	**0.968**	0.63	0.926	0.71
Recall	**0.996**	0.77	0.921	0.57	0.939	0.50
F1-Score	0.802	0.72	**0.944**	0.60	0.932	0.59
Porpoise Click						
Precision	0.868	0.68	**0.966**	0.67	0.925	0.69
Recall	0.985	0.73	0.957	0.53	**0.979**	0.48
F1-Score	0.922	0.71	**0.961**	0.59	0.951	0.57

(a) Multi-labels trained classifiers performances.



(b) Multi-classes trained classifiers performances.

Figure 6: Comparison of classifiers trained with multi-labels (top row) vs. multi-classes approaches(bottom row) before integration of attention masks. Values are normalized by the size of the test set and represent the percentage of well classified labels.