Multi-Modal Attention Framework for Underwater Bioacoustic Denoising and Recognition

Anonymous Author(s)
Affiliation
Address
`email`

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.

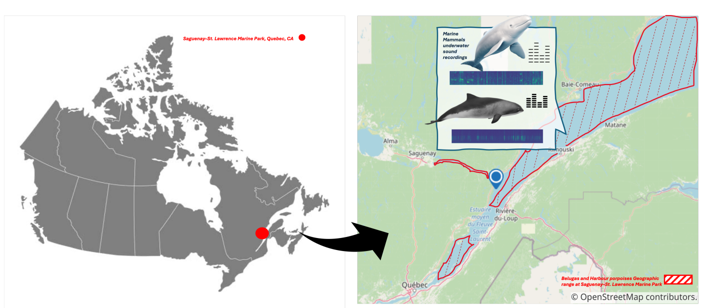


Figure 1: Saguenay–St. Lawrence Marine Park (SSLMP) representation.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

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-modal 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 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,

As a transfer learning strategy [14], ImageNet normalization was applied to all inputs, given that most models were pretrained on this dataset.

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.

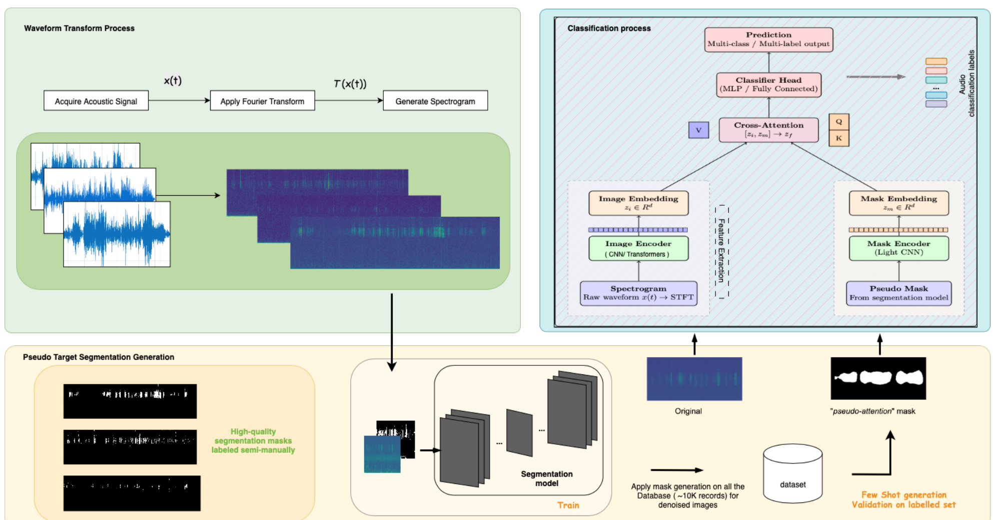


Figure 2: End-to-end framework for automatic denoising and classification from raw audio.

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-modal 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.

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.

Figure 3: Spectrogram (**left**), high-quality segmentation mask (**middle**), and generated pseudo-attention mask (**right**) for a recording of porpoise clicks.

Mask-guided multimodal 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 modality [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 (see fig. 3. Concretely, we design a multimodal 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 multimodal 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 multimodal 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

Model	Accuracy	F1 macro
ResNet50	0.588	0.562
ConvNeXt	0.625	0.591
ViT	0.788	0.787
Multimodal (Gen. masks)	0.837	0.816
Multimodal (HQ masks)	**0.897**	**0.890**

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

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.

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 modalities.

5 Discussion

While our framework demonstrates promising results, it inherits some limitations from the signal transformation choices. In particular, STFT can introduce resolution trade-offs and information loss, which may restrict the model's ability to fully capture the complexity of marine mammal vocalizations. Moreover, our study did not incorporate explicit uncertainty quantification, an aspect that is increasingly important for trustworthy machine learning in ecological monitoring. Future work will address these issues by exploring alternative time–frequency representations, improving attention mechanisms, and integrating methods to quantify predictive uncertainty, thus making the framework more robust and reliable for scientific and conservation-oriented applications.

6 Conclusion

We introduced a segmentation-guided multimodal framework that consistently improves recognition of marine mammal vocalizations under real-world noise and overlap. By fusing spectrogram and mask embeddings via mid-level cross-attention, the method produces reliable and interpretable presence signals that align with independent visual surveys. This establishes a principled route to scientific inference from raw acoustic signals, with immediate relevance to ecology and broader acoustic sensing problems. Overall, our results demonstrate that deep learning models can extract reliable presence signals that directly support species monitoring and conservation, illustrating how these 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.

7 Annexe

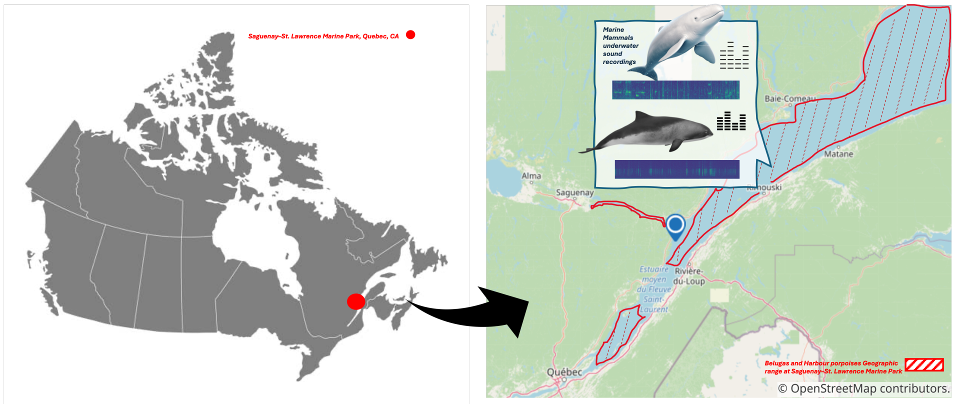


Figure 4: Saguenay–St. Lawrence Marine Park (SSLMP) representation.

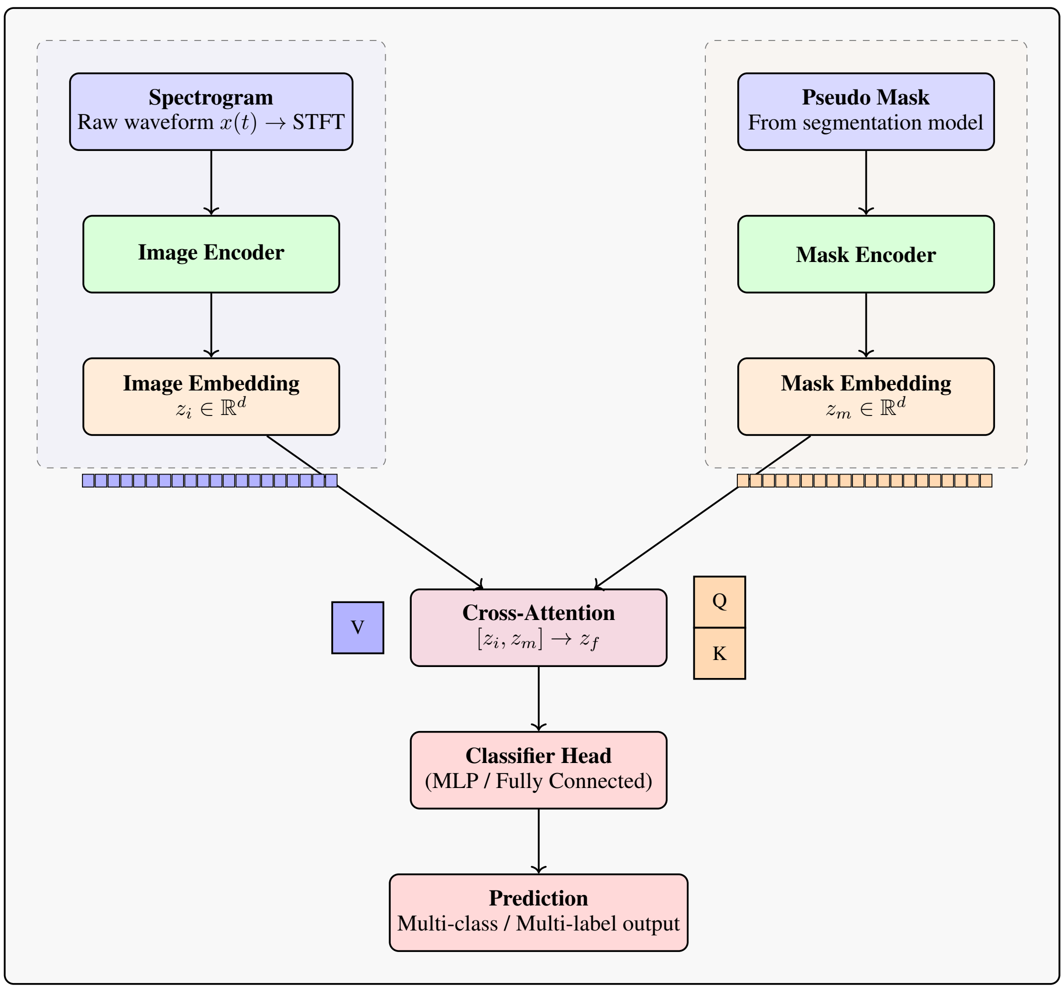
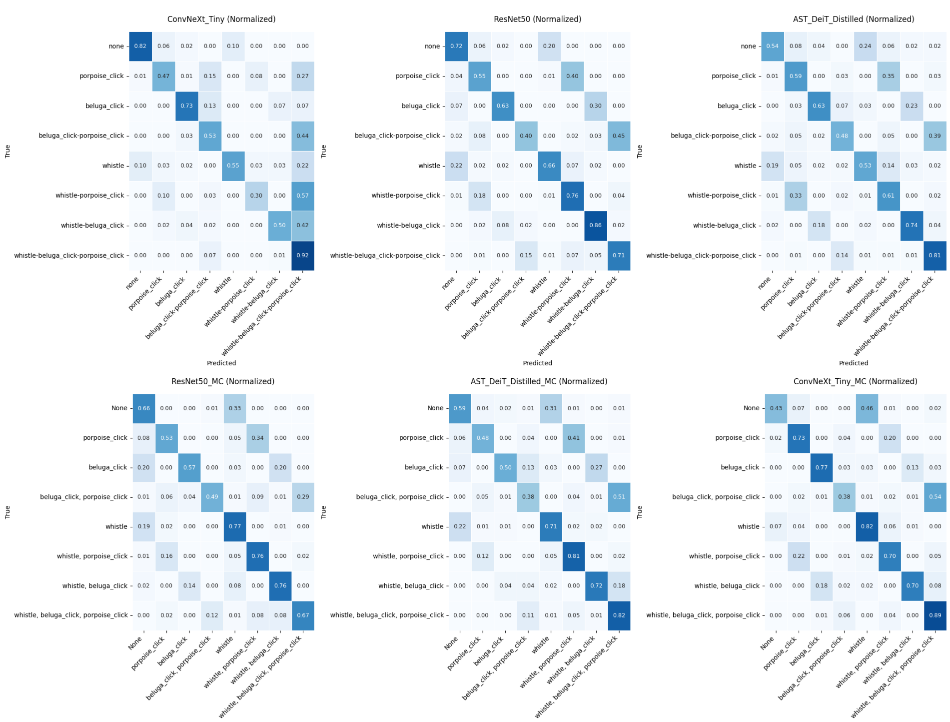


Figure 5: Architecture of the proposed model with two encoding branches and mid-fusion by cross-attention

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.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main contribution is the proposition of a framework, with details of its components provided in 2. We demonstrate the efficiency of the approach in Section 4.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations are discussed at the end of the paper and are the subject of ongoing and future research (see 6).

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: We justify our assumptions throughout the paper. For the machine learning results, we provide empirical evidence in addition to theoretical arguments to support our claims as comprehensively as possible.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: For reproducibility, we relied on existing baseline models, which are cited accordingly. For our own model, we provide a detailed description of the architecture in Section 3.1, ensuring that the framework can be replicated without difficulty.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We are making our best efforts to share the data and code, but they remain private for the moment. We plan to make them publicly available soon to contribute to the research community.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We explain the details of the training methodology in section 3.1, and the details of the datasets used for training and testing are provided in section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: The main objective of this paper is to introduce and validate a complete framework for improving the classification of underwater recordings. While the reported results are consistent and reproducible across several experimental settings, we did not include explicit error bars or confidence intervals. This choice was made to keep the focus on demonstrating the methodological contributions and their relative improvements over baseline approaches, rather than on an in-depth statistical uncertainty analysis. Nevertheless, the experiments were run under fixed and well-documented conditions (same datasets, train/test splits, and evaluation metrics), ensuring that the reported performance is reliable and can be independently reproduced. Incorporating a more detailed uncertainty quantification is left as a direction for future work.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: When presenting results for the various models, we describe the computational resources used and provide an estimate of the running time (see Table 2).

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: We consider ethics a fundamental aspect of our research. This paper is written in accordance with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The work discusses the impact underwater noise analysis for marine mammals monitoring as explained in the paper but more specifically in the abstract and the introduction.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: the paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Please see refrences.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: the paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: the paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

Justification: the paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: the core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.