COMPARISON OF VISION-BASED OBJECT DETECTORS AND AN AUDIO TRANSFORMER FOR WHALE SOUND EVENT DETECTION

Technical Report

Maaz Saeed[1], *Amir Latifi Bidarouni*[1], *Hanna Lukashevich*[1], *Jakob Abeßer*[2,1],

[1] Semantic Media Technologies, Fraunhofer IDMT, Ilmenau, Germany,
{maaz.saeed, amir.latifi.bidarouni, hanna.lukashevich}@idmt.fraunhofer.de
[2] Computational Humanities, University of Bamberg, Bamberg, Germany,
jakob.abesser@uni-bamberg.de

ABSTRACT

In this report, we address the automated detection of whale vocalizations in low-sampling-rate passive acoustic monitoring (PAM) recordings as part of BioDCASE Challenge task 2. We study two spectrogram-based object detectors from the computer vision domain, YOLOv11 and RT-DETR, alongside an audio-native transformer model, BEATs, fine-tuned on upsampled waveforms. Using an event-level evaluation based on 1D intersection-over-union (IoU) along the temporal axis, we observe that BEATs models deliver the highest performance among the base models (F1 score = 0.655), suggesting that large-scale audio pre-training transfers well even to narrowband, low-frequency whale vocalizations. RT-DETR clearly surpasses YOLO (Base.) in performance (F1 score of 0.517 vs. 0.430), and a detailed analysis of model-agnostic post-processing reveals that adding a simple non-maximum supression (NMS) step is essential for boosting instance-level performance, increasing RT-DETR's F1 score to 0.679. These results demonstrate that (i) self-supervised transformer models pre-trained on large-scale audio datasets already constitute strong baselines for marine bioacoustic event detection and (ii) transformer-based detectors with carefully tuned NMS remain competitive for precise temporal localization, offering a complementary pathway when frame-level spectrograms and high-resolution instance-level annotations are available.

***Index Terms*—** SED, object detection, transformer, non-maximum supression

1. INTRODUCTION

Recent bioacoustic and ecoacoustic studies use passive acoustic monitoring (PAM) to characterize underwater soundscapes, track vocal marine taxa, and assess the impacts of anthropogenic noise on marine organisms over large spatial and temporal scales [1, 2]. For instance, PAM studies of baleen whales have shown that temporal and individual variation in call production reflects movement patterns, behavioural states, and seasonal habitat use, underscoring the value of vocal behaviour as an ecological indicator [3]. These analyses build upon curated reference repositories, including the Watkins Marine Mammal Sound Database [4] and a recent initiative towards a global library of underwater biological sounds [5]. The ongoing deployment of hydrophones generates such large quantities of underwater sound recordings that manual analysis becomes impractical.

Methodological reviews document a shift from traditional energy-based and spectrogram-based detectors to machine-learning approaches and deep neural networks that enable robust detection and classification of calls from marine mammals and fish in complex and noisy environments [6, 7, 8]. Transformer-based audio models such as Audio Representation Learning with Teacher-Student Transformer (ATST) [9] and Bidirectional Encoder representation from Audio Transformers (BEATs) [10] have led to substantial performance gains on general sound event detection (SED) tasks. Trained in a self-supervised manner on large-scale audio datasets, these architectures can be efficiently fine-tuned for downstream problems, which makes them particularly attractive for whale-call detection where labeled data are scarce and costly to obtain. Extending this paradigm, recent audio–language foundation models for bioacoustics, including self-supervised transformers tailored to rare-event detection directly from time-domain audio waveforms, have been proposed as scalable representation-learning frameworks with strong potential to generalize across diverse bioacoustic monitoring applications [11, 12].

In parallel, object detection architectures from computer vision—most notably the You Only Look Once (YOLO) [13] family of models and recent transformer-based detectors such as Real-Time DEtection TRansformer (RT-DETR) [14] have achieved state-of-the-art performance in localizing and classifying objects in images. By treating time–frequency representations of audio signals, like spectrograms, as images and modeling acoustic events with time–frequency bounding boxes, these architectures can be adapted for use in SED. However, because these models are originally pre-trained on images instead of audio, using them in the acoustic domain requires training or fine-tuning them on spectrogram-based inputs.

The BioDCASE Challenge task 2 targets automatic detection of Antarctic blue and fin whale vocalizations in continuous underwater recordings from long-term deployments. The task is defined as supervised, multi-class, multi-label sound event detection with strong labels, where models must both classify call type and temporally localize events in long-term passive acoustic recordings. Evaluation uses a 1D intersection-over-union (IoU) measure along the temporal axis, and the 7 labeled call types are additionally combined into three biologically meaningful aggregated classes for scoring, in order to preserve robustness to sparse events and the highly variable acoustic conditions in Antarctica.

In this study, we address the challenge of detecting whale sound events in the BioDCASE Challenge task 2 framework by fine-tuning

object detection architectures, specifically YOLOv11 [15] and RT-DETR, and comparing with audio-specific pre-trained transformer model, BEATs. In this work we applied modern object detectors (e.g., YOLO, RT-DETR) operating on spectrograms, where whale calls are encoded as time–frequency bounding boxes that are directly compatible with the challenge's event-based evaluation and applied data augmentation. In addition, we adapt pre-trained BEATs to an event-detection setup to obtain accurate onset–offset estimates and compare these two paradigms—spectrogram-based object detection versus audio-transformer-based SED—on the BioDCASE Challenge task 2 dataset. Finally, we investigate post-processing strategies to further improve the original RT-DETR performance.

2. DATASET

The BioDCASE Challenge task 2 dataset consists of continuous underwater passive acoustic recordings from 11 site–year deployments around Antarctica collected between 2005 and 2017. All recordings are consistently resampled to 250 Hz and primarily capture whale calls in the 20–120 Hz range. They are annotated with onset and offset times and with frequency limits for seven call types—bma, bmb, bmz, bmd, bp20, bp20plus and bpd—emitted by Antarctic blue and fin whales. These call types are grouped into three classes, labeled bmabz, bp and d, which together account for only about 5 % of the total recording duration. The training set includes 6,004 recordings ranging from 4 min 59 s to 1 h in length, for a total of 1,292 h and 58,570 annotated events. The validation set comprises 587 one-hour recordings with 17,627 events in total, while the evaluation set contains 795 one-hour recordings whose annotations are withheld for benchmarking.

3. PROPOSED METHOD

3.1. Deep Learning Models

In this study, we fine-tune three deep learning models for SED. Two of these, YOLOv11 and RT-DETR, are object detectors originally pre-trained on large-scale natural image datasets and here repurposed by training on time–frequency audio representations for SED. The third model, BEATs, is a transformer-based audio representation model pre-trained in a self-supervised fashion on large-scale audio data and subsequently fine-tuned directly for the target SED task.

3.2. Feature Extraction Parameters

While BEATs operates on audio resampled from 250 Hz to 16 kHz, the object detection models use the original 250 Hz signals, which are directly converted into images without further resampling. For object detection models, we use an FFT size of 512, a window length of 256 samples and a hop size of 20 samples as the feature configuration, resulting in an 80 ms time resolution (frame size).

3.3. Pre-processing and Data Augmentation

Audio recordings are segmented into 30-s excerpts with 50% overlap, which are provided as input to BEATs. For downstream object detection with YOLO and RT-DETR, we generate spectrograms for each segment, then standardize them across the frequency axis via z-score normalization, and finally perform min–max scaling. Since these models require three-channel inputs, the single-channel spectrograms are replicated across the three channels to match the input format used during pre-training of the object detection models. Bounding boxes are defined in the spectrogram image space by mapping the horizontal axis to time and the vertical axis to frequency. Onset and offset times are converted to normalized coordinates relative to the segment duration, while the vertical extent corresponds to the relevant frequency range. As BEATs predicts only temporal boundaries, its outputs are likewise transformed into bounding boxes with the full vertical span of the spectrogram to represent the entire frequency range.

For the YOLOv11 model, the official Ultralytics configuration applies on-the-fly HSV color jitter, random horizontal flipping, translation, RandAugment, Random Erasing, Mosaic data augmentation that combine four images into a single training sample. For the RT-DETR model, we follow the default RT-DETR setup, which primarily employs random horizontal flipping, multi-scale random resizing of the input images, and random cropping of image–annotation pairs.

3.4. Fine-tuning and Post-processing

We conducted experiments with the BEATs, RT-DETR, and YOLOv11 models, systematically varying the use of data augmentation and the training batch size. For each model–batch-size configuration, we computed the corresponding Micro-averaged precision, recall, and F1 scores on the evaluation set. The resulting configurations and metrics are summarized in Table 1.

Decreasing the detection confidence threshold to an optimal range of 0.3-0.4 tends to increase the F1 score, as additional true positives are recovered. However, beyond this range, further reductions in the threshold lead to a decrease in F1 score due to a disproportionate increase in false positives. These false positives may arise from misclassified events, inaccurately localized bounding boxes, or multiple highly overlapping detections corresponding to a single ground-truth event, all of which are penalized by the evaluation metric. These observations indicate that confidence-threshold optimization and appropriate post-processing are critical for maximizing detector performance.

YOLO applies non-maximum supression as a default step to mitigate redundant overlapping bounding box detections. non-maximum supression retains only the highest-confidence prediction within each group of overlapping boxes. All other boxes whose intersection-over-union (IoU) with this retained box exceeds a predefined threshold are discarded. Applying NMS as a post-processing procedure to suppress overlapping bounding boxes detected by RT-DETR is illustrated in Fig. 1.

To exploit systematic differences in bounding box aspect ratios between true events and false positives, we introduce an aspect ratio filtering of predicted bounding boxes:

$$\text{Aspect Ratio: AR} = \frac{\text{bounding box width}}{\text{bounding box height}}, \tag{1}$$

The empirical aspect ratio distributions for the seven original call types and for the three grouped classes are shown in Fig. 2 and Fig. 3, respectively. In the grouped representation Fig. 3, the d and bp classes exhibit sharply peaked distributions at very small aspect ratios ($AR \lesssim 0.5$), indicating that almost all events in these classes correspond to vertically elongated bounding boxes. In contrast, the $bmabz$ class displays a much broader distribution spanning $AR \approx 0.5$–4, with a clear mode around $AR \approx 1$–2 and a long

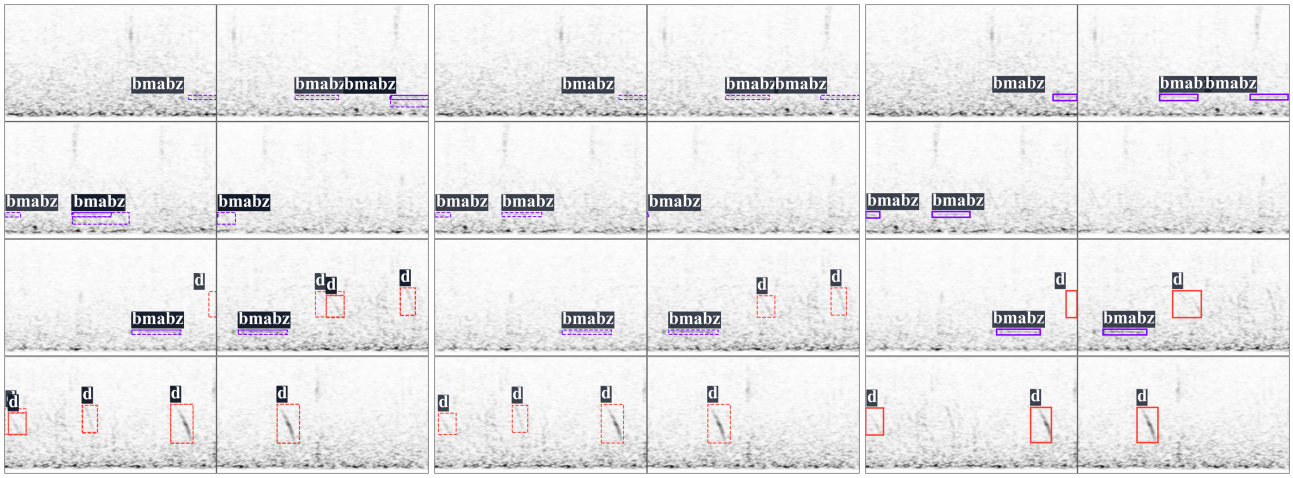

Figure 1: Whale call detections by the RT-DETR model shown without (left) and with (middle) NMS post-processing. Purple and red bounding boxes indicate bmabz and d call types, respectively. Ground-truth labels are shown for reference (right).

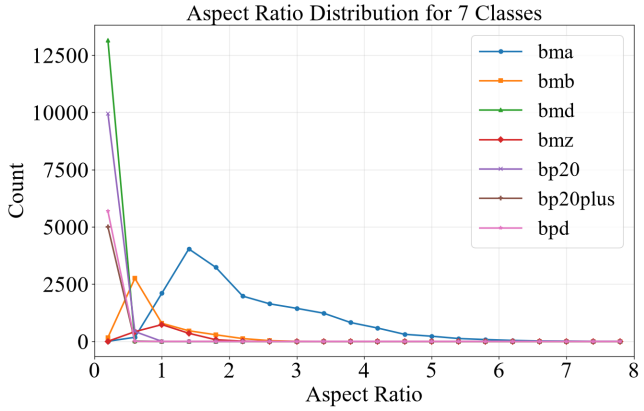

Figure 2: Bounding box aspect ratio distribution across 7 call types

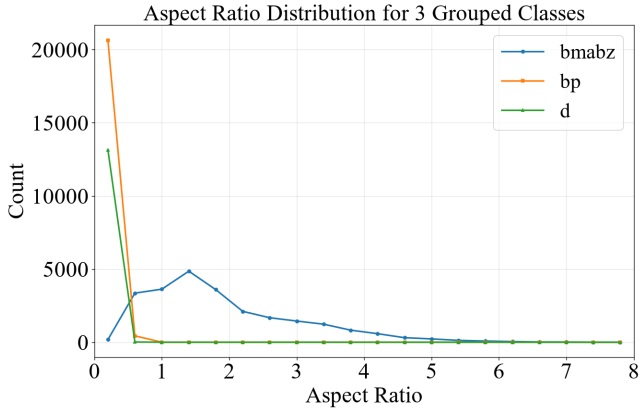

Figure 3: Bounding box aspect ratio distribution across 3 aggregated classes

tail towards larger aspect ratios. This pattern reflects the fact that *bmabz* events tend to be temporally extended and therefore appear as horizontally elongated or semi-square boxes. Consistent trends are already visible in the seven-class distributions (Fig. 2), where the various *bp*-type calls are concentrated at very low aspect ratios, whereas *bm*-type calls occupy a wider range with a pronounced bias towards larger, more horizontal boxes.

Based on the class-specific aspect ratio distributions observed in the training data, we define an expected AR range for each class and use it to apply aspect ratio filtering as a post-processing step during inference. A predicted event is retained only if its aspect ratio falls within the expected range for its predicted class; otherwise, the prediction is discarded as likely incorrect. This class-dependent aspect ratio filtering provides an inexpensive geometric consistency check that helps suppress false positives.

In a qualitative analysis of the detector outputs, we observed that single events are often fragmented into multiple, closely spaced detections. To mitigate this, we apply a box-merging post-processing step. Two consecutive predicted events of the same class are merged into a single event if (i) the temporal gap between them is smaller than a predefined threshold and (ii) their frequency extents sufficiently overlap. This criterion also implicitly handles overlapping duplicate detections, since overlapping boxes have zero temporal gap. The merged event inherits the earliest start time and the latest end time among the constituent boxes. This procedure reduces duplicate and fragmented predictions, which is crucial because the evaluation penalizes multiple detections corresponding to the same ground-truth event.

Fig. 4 illustrates this process: the raw predictions (orange dashed boxes) split a single `bmabz` call into two fragments, which are merged into a single event (purple dashed box) spanning the full event duration.

Since RT-DETR does not natively include such post-processing, we systematically assess the impact of combining confidence-threshold optimization with non-maximum supression, aspect ratio filtering, and bounding box merging on its performance. The performance of these data-driven schemes is summarized in Table 2.

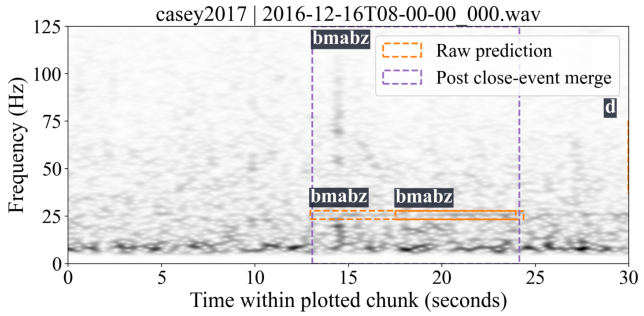

Figure 4: Box merging of two fragmented detections of a single `bmabz` event.

4. EVALUATION

Table 1 summarizes the Micro-averaged precision, recall, and F1 scores for all model–batch-size configurations. The YOLO (Base.) model, fine-tuned without data augmentation YOLO (Base.), attains a relatively low F1 score of 0.430, driven by poor precission (0.320). Introducing augmentation YOLO (Aug.) raises recall considerably to 0.767, but at the cost of a reduction in precision, which produces an even lower F1 score (0.356) and suggests that this computer-vision pre-trained model generalizes poorly to unseen audio events. For RT-DETR, reducing the batch size from 32 to 16 does not produce a notable change in F1 score, but its F1 score of 0.517 still represents roughly a 20 % relative improvement over the YOLO (Base.). Among all architectures, BEATs achieves the highest precision (0.619) and the best overall F1 score (0.655), indicating the strongest generalizability to the downstream audio detection task.

Model	Batch size	Recall	Precision	F1 score
YOLO (Base.)	–	0.670	0.320	0.430
YOLO (Aug.)	32	0.767	0.232	0.356
RT-DETR	32	0.699	0.403	0.511
RT-DETR	30	0.700	0.394	0.503
RT-DETR	16	0.665	0.423	0.517
BEATs	8	0.696	0.619	**0.655**

Table 1: Micro-averaged precision, recall, and F1 scores for different models and batch sizes.

The BEATs model outperforms all spectrogram-based object detectors, which we attribute to its audio-native pre-training on large-scale corpora and its ability to learn modulation patterns and long-range temporal structure that transfer even to narrowband, low-frequency data. Although the whale recordings are sampled at 250 Hz and upsampled to 16 kHz for compatibility, this preserves the salient low-frequency energy and the smooth, contour-like time–frequency trajectories of whale calls, which BEATs' self-supervised representations can exploit despite the limited bandwidth. In contrast, YOLOv11 and RT-DETR rely on backbones pre-trained on natural images, making them better suited to static spatial textures than to continuous spectro-temporal contours. RT-DETR's transformer layer improves sequential event modeling relative to YOLOv11, but its vision-centric pre-training still limits its ability to learn robust acoustic features from scarce labels. Overall, the audio-specific inductive biases and large-scale pre-training of BEATs align more closely with the structure and frequency range of whale vocalizations, yielding the strongest generalization and highest F1 score among the evaluated models.

Post-processing scheme	Micro-averaged F1 score
–	0.517
NMS	**0.679**
aspect ratio filtering	0.502
Box merging	0.666
NMS + filters + merging	0.635

Table 2: Micro-averaged F1 scores for different post-processing schemes.

Table 2 reports the effect of different post-processing schemes on the Micro-averaged F1 score of the best-performing base model (RT-DETR). Without any post-processing ("–"), the model attains a Micro-averaged F1 score of 0.517. Applying non-maximum supression ("NMS") alone yields the highest F1 score (0.679), indicating that suppressing redundant, overlapping detections is the most effective strategy. This finding is consistent with our event-level evaluation, which penalizes multiple detections of the same ground-truth event as additional false positives. By collapsing overlapping boxes, NMS aligns the predicted instance set more closely with the evaluation protocol and thus directly boosts F1 score. Box merging also improves performance (F1 score = 0.666), but remains slightly below NMS, likely because overly aggressive merging can fuse distinct nearby events. aspect ratio filtering alone degrades performance (F1 score = 0.502), and the combined "NMS + filters + merging" scheme (F1 score = 0.635) underperforms NMS, suggesting that additional heuristic filters beyond a well-tuned NMS provide limited benefit and can even remove valid detections.

5. CONCLUSION

In this work, we compared spectrogram-based object detectors (YOLOv11, RT-DETR) with an audio-native transformer model (BEATs) for whale-call sound event detection. Across three base configurations, BEATs achieved the highest F1 score (0.655), highlighting the advantage of large-scale, audio-specific pre-training for low-frequency marine bioacoustics. Among the object detectors, RT-DETR substantially outperformed YOLO (Base.) (0.517 vs. 0.430 F1 score), and a simple, task-aligned post-processing scheme based on non-maximum supression further boosted RT-DETR to 0.679 F1 score, surpassing BEATs by a small margin. The post-processing analysis shows that carefully tuned NMS is more beneficial than more complex heuristic filtering or box-merging strategies. Overall, these results suggest that audio foundation models already provide strong baselines in data-scarce marine settings, while transformer-based detectors with principled post-processing remain competitive and complementary tools for fine-grained temporal localization of underwater acoustic events.

6. REFERENCES

[1] C. Peng, X. Zhao, and G. Liu, "Noise in the sea and its impacts on marine organisms," *International Journal of Environmental Research and Public Health*, vol. 12, no. 10, pp. 12 304–12 323, 2015. [Online]. Available: https://www.mdpi.com/1660-4601/12/10/12304

[2] M. Solé, K. Kaifu, T. A. Mooney, S. L. Nedelec, F. Olivier, A. N. Radford, M. Vazzana, M. A. Wale, J. M. Semmens, S. D. Simpson, G. Buscaino, A. Hawkins, N. Aguilar de Soto, T. Akamatsu, L. Chauvaud, R. D. Day, Q. Fitzgibbon, R. D. McCauley, and M. André, "Marine invertebrates and noise," *Frontiers in Marine Science*, vol. Volume 10 - 2023, 2023. [Online]. Available: https://www.frontiersin.org/journals/marine-science/articles/10.3389/fmars.2023.1129057

[3] T. A. Webster, S. M. Van Parijs, W. J. Rayment, and S. M. Dawson, "Temporal variation in the vocal behaviour of southern right whales in the auckland islands, new zealand," *Royal Society Open Science*, vol. 6, no. 3, p. 181487, 03 2019.

[4] L. Sayigh, M. A. Daher, J. Allen, H. Gordon, K. Joyce, C. Stuhlmann, and P. Tyack, "The watkins marine mammal sound database: An online, freely accessible resource," *Proceedings of Meetings on Acoustics*, vol. 27, no. 1, pp. 1–9, 2016.

[5] M. J. G. Parsons, T.-H. Lin, T. A. Mooney, C. Erbe, F. Juanes, M. Lammers, S. Li, S. Linke, A. Looby, S. L. Nedelec, I. Van Opzeeland, C. Radford, A. N. Rice, L. Sayigh, J. Stanley, E. Urban, and L. Di Iorio, "Sounding the call for a global library of underwater biological sounds," *Frontiers in Ecology and Evolution*, vol. Volume 10 - 2022, 2022. [Online]. Available: https://www.frontiersin.org/journals/ecology-and-evolution/articles/10.3389/fevo.2022.810156

[6] M. Bittle and A. J. Duncan, "A review of current marine mammal detection and classification algorithms for use in automated passive acoustic monitoring," 2013. [Online]. Available: https://api.semanticscholar.org/CorpusID:11858992

[7] M. Zhong, M. Castellote, R. Dodhia, J. Lavista Ferres, M. Keogh, and A. Brewer, "Beluga whale acoustic signal classification using deep learning neural network models," *The Journal of the Acoustical Society of America*, vol. 147, no. 3, pp. 1834–1841, 03 2020. [Online]. Available: https://doi.org/10.1121/10.0000921

[8] M. Zhong, M. Torterotot, T. A. Branch, K. M. Stafford, *et al.*, "Detecting, classifying, and counting blue whale calls with siamese neural networks," *The Journal of the Acoustical Society of America*, vol. 149, no. 5, pp. 3086–3094, 2021.

[9] X. Li, N. Shao, and X. Li, "Self-supervised audio teacher-student transformer for both clip-level and frame-level tasks," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 32, p. 1336–1351, Jan. 2024. [Online]. Available: https://doi.org/10.1109/TASLP.2024.3352248

[10] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, W. Che, X. Yu, and F. Wei, "BEATs: Audio pre-training with acoustic tokenizers," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 23–29 Jul 2023, pp. 5178–5193. [Online]. Available: https://proceedings.mlr.press/v202/chen23ag.html

[11] J. C. Schäfer-Zimmermann, V. Demartsev, B. Averly, K. L. Dhanjal-Adams, M. Duteil, G. Gall, M. Faiß, L. Johnson-Ulrich, D. Stowell, M. B. Manser, M. A. Roch, and A. Strandburg-Peshkin, "animal2vec and meerkat: A self-supervised transformer for rare-event raw audio input and a large-scale reference dataset for bioacoustics," *Methods in Ecology and Evolution*, vol. 17, no. 3, p. 875–888, Dec. 2025. [Online]. Available: http://dx.doi.org/10.1111/2041-210x.70218

[12] D. Robinson, M. Miron, M. Hagiwara, and O. Pietquin, "NatureLM-audio: an audio-language foundation model for bioacoustics," 2024.

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

[14] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, and J. Chen, "Detrs beat yolos on real-time object detection," 2024. [Online]. Available: https://arxiv.org/abs/2304.08069

[15] G. Jocher and J. Qiu, "Ultralytics yolo11," 2024. [Online]. Available: https://github.com/ultralytics/ultralytics