# ARIA: A Hybrid BirdNET-PERCH Framework for Inventory-Driven Bird Species Recognition in Zoo Aviaries

Emre Argın, Bernardo Amado Costa, Aki Härmä, Aysenur Arslan-Dogan
DACS, Maastricht University
Maastricht, Netherlands
{aki.harma,aysenur.arslan-dogan}@maastrichtuniversity.nl

***Abstract*—Zoo aviary soundscape monitoring is an inventory recovery task: each deployment has a known set of expected taxa and operational value depends on reliably detecting that inventory under acoustically challenging conditions. Aviaries introduce reverberation, dense multi-species overlap, and taxonomic composition that diverges from public training corpora, causing general-purpose classifiers to fail structurally. We present ARIA (Acoustic Recognition for Inventories of Aviaries), a hybrid detection framework designed specifically for inventory-driven monitoring in captive settings. ARIA integrates head-only retrained BirdNET for domain adaptation, PERCH v2 for broad species coverage and acoustic embeddings, and a learned fusion mediator that combines multi-label predictions with complementary representations while mitigating confidence artifacts induced by label-set constraints. Deployment-aware filtering and weighted ensemble inference ensure biologically plausible outputs. Evaluation combines segment-level manual annotation for direct agreement assessment with full-corpus operational metrics across 16 zoo deployments. ARIA substantially reduces empty-prediction rates on confirmed vocalizations and improves inventory coverage relative to constrained default BirdNET, enabling scalable acoustic monitoring for avian welfare and collection management.**

***Index Terms*—Bioacoustic monitoring, BirdNET, PERCH, domain adaptation, model fusion, bird sound classification, zoo soundscapes**

## I. Introduction

Automated bioacoustic monitoring has become essential for continuous, non-invasive tracking of animal populations, with deep learning classifiers enabling large-scale soundscape analysis [1]. BirdNET [2], trained on millions of wild bird recordings, exemplifies this capability. Yet specific local environments present distinct challenges that expose fundamental limitations in general-purpose classifiers. In this paper, we focus on zoo aviaries, where species composition is known in advance–a constrained inventory recovery problem where operational value depends on confirming that known taxa vocalize under acoustically challenging conditions.

Captive soundscapes systematically differ from wild settings [3]. Enclosures introduce reverberation; concurrent vocalizations create dense temporal overlap; institutional collections include taxa underrepresented in public archives. General-purpose models respond to this domain mismatch with structural detection failures: when constrained to site-specific species lists and filtered by confidence thresholds, default BirdNET returns empty predictions for most audio segments, missing audible target species entirely. PERCH v2 [4] offers broader taxonomic coverage but introduces complementary failure modes: its softmax activation precludes multi-label detection of overlapping calls and constraining logits to small deployment-specific label sets artificially inflates confidence regardless of acoustic evidence.

We introduce ARIA, a hybrid detection framework addressing these complementary failures through three components. Head-only retrained BirdNET adapts the pretrained feature extractor to zoo-relevant taxa, expanding vocabulary coverage while preserving multi-label capability essential for overlapping vocalizations. PERCH v2 provides complementary taxonomic coverage and acoustic embeddings carrying discriminative information when BirdNET yields low-confidence or null predictions. A learned fusion network combines BirdNET's multi-label outputs with PERCH v2 representations through separate projection heads, acting as a mediator that recovers detections in BirdNET-null segments while dampening confidence distortions from label-set constraints. Deployment-aware species filtering excludes biologically impossible taxa and weighted consensus inference aggregates evidence, weighting BirdNET as the primary multi-label detector, discounting PERCH v2 due to its single-label constraint and boosting predictions where models agree. Beyond its zoo-monitoring application, the learned fusion mediator constitutes a reusable mechanism for mitigating softmax-induced confidence inflation when foundation classifiers are constrained to small, domain-specific label sets — a failure mode likely to arise in any constrained-inventory recognition task.

We evaluate ARIA on a test corpus collected from 16 zoo enclosures comprising 999,177 audio segments and 97 species—entirely held out from any model training. Evaluation combines segment-level manual annotation on a stratified sample for direct agreement assessment with full-corpus operational metrics across all deployments. Results demonstrate that domain adaptation alone improves segment-level agreement from below 50% to approximately 80%, while hybridization provides incremental gains on segments where BirdNET produces no output. These improvements enable scalable acoustic monitoring that can inform welfare assessment and collection management decisions.

Fig. 1 provides a high-level overview of the offline training and model-building pipeline described in Sections II–V.

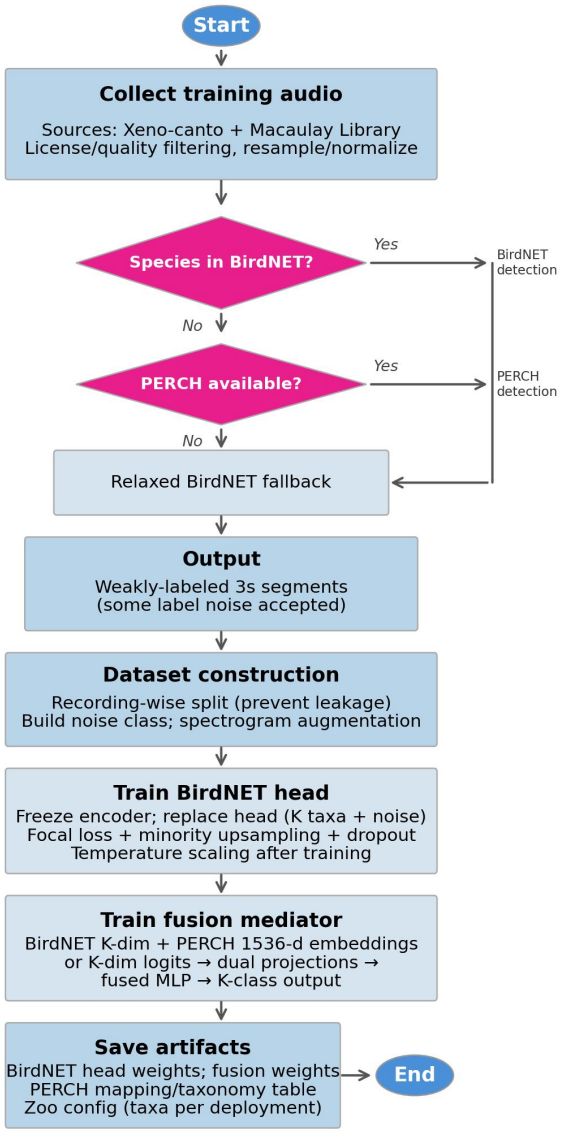



Fig. 1: Offline training and model-building pipeline (overview).

## II. Training Data Acquisition and Segmentation

Training audio was obtained from Xeno-canto [5] through an automated pipeline that queried the API for each target species and downloaded matching recordings, and from Cornell Lab of Ornithology | Macaulay Library [6] via special request. Recordings with incompatible licenses or corrupted content were excluded, and all valid audio was normalized to a fixed sampling rate.

Since Xeno-canto and Macaulay Library provide only file-level annotations for recordings that are typically several minutes long, we segmented each file into 3 s clips using automated vocal-event detection. The segmentation strategy was adapted to each species' representation in BirdNET: for taxa already present in BirdNET's training set, we applied species-constrained detection; for species absent from BirdNET, we used PERCH v2; and when neither detector produced results, we applied relaxed general-purpose BirdNET as a fallback. However, because source recordings lack precise temporal annotations and automated detection is imperfect, extracted segments may include mixed-species audio, low signal-to-noise vocalizations, or intervals without the target species. This is an accepted trade-off for achieving the scale required to cover diverse zoo taxa.

## III. Data Preprocessing and Augmentation

The classifier operates on fixed three-second audio segments at 48 kHz sampling rate, extracted from field recordings via automated segmentation. To construct the training set, we enforced minimum and maximum sample counts per species, removing underrepresented species with insufficient data while capping highly represented species to prevent dominance during optimization.

**Recording-wise data splits.** To prevent temporal leakage, we grouped segments by source recording and assigned entire recordings exclusively to either training or test splits. This ensures that no recording contributes segments to both sets, yielding realistic generalization estimates despite potential label noise and temporal correlation within recordings.

**Negative class construction.** Following BirdNET conventions, we introduced a dedicated noise class to provide explicit negatives. Candidate noise segments were sampled from recordings that did not contribute any positive examples, preventing audio from appearing as both signal and noise. Sliding windows with short steps generated candidate intervals; subsampling with per-recording caps promoted acoustic diversity. The resulting noise class captures microphone self-noise, wind, distant human activity and other non-focal events.

**Spectrogram augmentation.** For training classes below a minimum count threshold, we applied mel-power spectrogram augmentation to improve class balance and acoustic robustness, following the augmentation families of Lauha et al. [7]. Three-second waveforms were converted to power mel spectrograms and subjected to stochastic perturbations: frequency and time stretching via crop-and-interpolate; frequency/time shifts with edge shuffling; power-law contrast warping; and Poisson-distributed frequency or time masking. Noise mixing used noise segments at random mixing ratios. Augmented spectrograms were inverted to waveforms and RMS-normalized to match original loudness. A fixed fraction of samples remained unaugmented to anchor the model to real distributions. Per-source diversity caps prevented overfitting to individual source recordings while meeting target class sizes.

We evaluated ECOGEN [8], a VAE-based generative model for synthetic bird vocalizations, but found it degraded validation performance due to spectrogram parameter mismatches (FFT window size, sample rate) and overly aggressive synthetic-to-real ratios that introduced artifacts. We retained the spectrogram-

based approach, which operates directly on real recordings and preserves empirical acoustic structure.

## IV. BirdNET-based Classifier Architecture and Training

For zoo monitoring applications, default BirdNET's vocabulary exhibits limited alignment with institutional inventories [9]: key captive taxa are often absent while thousands of wild species remain. To address coverage gaps in our deployment, we retrained BirdNET [2] through head-only fine-tuning [10], [7]. We froze the pretrained convolutional feature extractor and replaced the original 6,500-class output layer with a new classification head for our target species plus a non-event noise class. Input spectrograms are processed by the frozen encoder; only the final dense layers are trainable.

Despite augmentation, the training set remained substantially imbalanced across classes. Under such imbalance, standard cross-entropy training can yield poorly calibrated confidence, with rarer species tending to receive more conservative scores that may hinder detection at fixed operational thresholds even when class boundaries are learnable. We addressed this through complementary training-time interventions.

**Focal loss.** We adopted focal loss [11], $\mathrm{FL}(p_t) = -\alpha(1 - p_t)^\gamma \log(p_t)$, which down-weights high-confidence predictions via the modulating factor $(1 - p_t)^\gamma$, redistributing gradient capacity from well-learned common species toward challenging examples. We fixed $\alpha = 0.25$ per multi-class conventions and tuned $\gamma$ to balance rare-species recall against calibration quality. Higher $\gamma$ suppresses well-learned examples more aggressively but can degrade confidence estimates.

**Minority upsampling.** Linear upsampling increased sampling probability for minority classes proportional to their deficit from uniform distribution, ensuring rare species appear in batches more frequently than raw frequencies suggest and preventing gradient-free epochs for underrepresented classes. The upsampling ratio controls rebalancing strength.

**Regularization.** Dropout ($p = 0.20$) prevented overfitting to synthetic augmentations. We evaluated mixup [12] but observed anomalously high multi-species co-prediction rates, likely because soft label targets hindered learning of sharp boundaries between acoustically similar species. Label smoothing triggered numerical instability when combined with focal loss in BirdNET's multi-label framework. Both were excluded.

**Training configuration.** Systematic validation experiments revealed a fundamental trade-off: configurations emphasizing minority classes (high $\gamma$, aggressive upsampling) improved rare-species recall by 15–20 percentage points but degraded calibration, producing saturated scores for common species and paradoxically underconfident predictions for rare species despite improved detection. Conservative configurations maintained better calibration but failed to achieve adequate rare-species recall. We selected $\gamma = 1.5$, upsampling ratio $= 0.60$, learning rate $= 6 \times 10^{-5}$, dropout $= 0.20$, and batch size $= 112$ to balance these objectives. Training proceeded for up to 200 epochs with early stopping on validation loss. Post-training, we applied temperature scaling [13] to the logit layer for improved confidence calibration without altering prediction rankings.

## V. ARIA System

While the custom BirdNET classifier described in Section IV provides strong performance on well-represented species, two limitations motivated the development of a hybrid architecture. First, some target species remained absent from BirdNET's training set despite our data collection efforts. Second, the challenging acoustic characteristics of zoo environments such as closed enclosures with reverberation, overlapping vocalizations, and variable recording conditions, caused BirdNET to fail silently on a substantial fraction of audio segments. To address these gaps, we integrated PERCH v2 [4], Google's second-generation bird vocalization classifier, as a complementary model within a fusion architecture. Combining complementary models has shown promise for improving species identification in challenging soundscapes[14], though prior work has focused on sequential pipelines rather than the parallel fusion with learned mediation adopted here.

### A. Complementary Model Strategy

BirdNET's sigmoid activation enables true multi-label prediction: each species probability is computed independently, allowing simultaneous detection of multiple vocalizing birds within a single segment. This capability is essential in zoo aviaries where overlapping calls are common. PERCH v2, by contrast, employs softmax activation, constraining outputs to sum to unity and yielding only the single most dominant species per segment. Despite this limitation, PERCH v2 offers two advantages that justify its inclusion.

First, PERCH v2's default model covers approximately 15,000 species, substantially exceeding BirdNET's vocabulary and encompassing all our target taxa without retraining. Second, PERCH v2 produces 1536-dimensional embeddings that encode rich acoustic structure, potentially capturing discriminative features in segments where BirdNET's confidence is low.

However, PERCH v2 cannot serve as a standalone classifier for our application. Beyond the single-label limitation, constraining PERCH v2's logits to a small target set introduces an artificial confidence inflation: when softmax is applied over only 87 filtered logits (rather than 15,000), probability mass is forced to distribute among those species regardless of whether any are actually present. This effect is most severe when few species are permitted and can produce spuriously high-confidence predictions. We therefore position PERCH v2 as a complementary model whose outputs inform, but do not dominate final predictions.

### B. Fusion Model Architecture

To combine BirdNET and PERCH v2 outputs, we designed a fusion neural network that learns to weight their complementary representations. Rather than naive concatenation, the architecture employs separate projection heads for each input stream, allowing specialized transformations before fusion.

**Dual-head projection.** BirdNET's 87-dimensional temperature-scaled probabilities pass through a projection block (Linear $\rightarrow$ BatchNorm $\rightarrow$ ReLU $\rightarrow$ Dropout) yielding 512-dimensional features. PERCH v2 inputs (either 1536-dimensional embeddings or 87-dimensional filtered logits, depending on operational mode) undergo an analogous projection to 512 dimensions. Separating projections prevents the network from conflating semantically distinct representations: BirdNET outputs are calibrated per-species confidence estimates from a multi-label classifier, whereas PERCH v2 outputs encode acoustic feature distributions.

**Fusion layers.** The concatenated 1024-dimensional representation passes through progressively narrowing fully connected layers ($1024 \rightarrow 512 \rightarrow 256 \rightarrow 87$), each followed by batch normalization, ReLU activation, and dropout ($p = 0.3$). The final layer produces 87 logits, converted to probabilities via softmax during inference.

**Role of the fusion model.** The fusion network serves as a mediator rather than a primary classifier. It evaluates predictions from both upstream models, identifies agreement, and produces a balanced output that can recover species missed by BirdNET alone or resolve ambiguous segments. The fusion component is designed as a *single-label* classifier: restricting it to a single dominant hypothesis avoids additional competition and thresholding complexity in the presence of BirdNET's multi-label outputs. In ARIA, multi-label detection is therefore handled exclusively by the retrained BirdNET component, while fusion contributes complementary single-label prediction that stabilizes decisions and provides candidate support in ambiguous or BirdNET-null windows. Because fusion training leverages PERCH v2's broader species coverage, ARIA implicitly benefits from acoustic knowledge beyond BirdNET's original training distribution. The fusion network was trained exclusively on Xeno-canto and Macaulay Library audio (Section II) and no zoo deployment recordings were used, ensuring evaluation reflects generalization without data leakage. During fusion training, both the retrained BirdNET classifier and PERCH v2 encoder remain frozen; only the fusion network weights are optimized. Training is supervised by the species labels assigned during automated segmentation (Section II), with the fusion model learning to predict the correct species from the concatenated BirdNET probability and PERCH v2 representation inputs.

### C. Constrained and Unconstrained Operating Modes

The system supports two modes distinguished by species coverage and input dimensionality.

**Unconstrained mode.** PERCH v2 outputs are not filtered; the full 1536-dimensional embedding is passed to the fusion model. This preserves maximum acoustic information and avoids logit-boost artifacts, at the cost of a larger fusion network (hidden dimensions [1024, 512, 256]), longer training, and irrelevant bird species detection.

**Constrained mode.** PERCH v2 logits are extracted and filtered to the 87 target-species indices via a precomputed mapping file that aligns BirdNET and PERCH v2 taxonomies. The resulting 87-dimensional vector replaces the embedding, reducing combined input dimensionality from 1629 to 186 (86% reduction) and enabling a smaller fusion network ([256, 128, 64]). While computationally efficient, this mode inherits the logit-boost risk when the allowed species set is small. During inference time, ARIA uses this mode.

In both modes, the same 1536-dimensional acoustic representation is computed internally by PERCH v2; the distinction lies in what is forwarded to fusion. When filtering is unnecessary, embeddings suffice; when targeting specific species, explicit logit extraction is required to select relevant indices.

### D. Zoo-Specific Biological Constraints

A critical feature of the ARIA is dynamic, deployment-aware species filtering. A configuration file specifies expected species per zoo and enclosure; the system parses dataset directory names (e.g., `zoo_city_aviary_x_data`) to automatically load the corresponding species list without manual intervention.

**Pre-filtering** is applied to PERCH v2 and the fusion model: logits are constrained during extraction, limiting predictions to biologically plausible taxa. **Post-filtering** is applied to BirdNET at inference time; because BirdNET assigns species confidences independently, pre-filtering would not alter its internal computations, only the reported outputs. Both filters exclude species that cannot be present at the recording location, substantially reducing false positives. Wild species common across deployments remain permitted regardless of enclosure.

### E. Ensemble Inference Strategy

Final predictions aggregate outputs from all three models via weighted voting.

**Weighted score aggregation.** Each model contributes to a species' vote proportionally to an assigned weight multiplied by its confidence: BirdNET receives weight 1.0 (reflecting its primary role and multi-label capability), fusion 0.8, and PERCH v2 0.6 (discounted due to logit-boost concerns and single-label limitation).

**Consensus boost.** Species predicted by two or more models receive an additive bonus (default 0.5), rewarding cross-model agreement and increasing robustness to individual-model errors.

**Name normalization.** Taxonomic naming conventions differ across models (e.g., "*Turdus merula*_Common Blackbird" vs. "Common Blackbird"); the voting system normalizes species names to prevent duplicate entries and ensure correct aggregation.

This ensemble strategy is particularly valuable at decision boundaries: when models disagree, voting weights toward consensus; when one model produces a high-confidence prediction unsupported by others, that prediction can still dominate if sufficiently strong. The result is a hybrid system that leverages BirdNET's multi-label strength, PERCH v2's acoustic richness, and fusion's learned weighting to achieve more robust detection across diverse zoo soundscapes than any single model alone.

Fig. 2 summarizes the online inference workflow executed per zoo deployment; this is the operational procedure used to generate all retained detections analyzed in Sections VI–VIII.

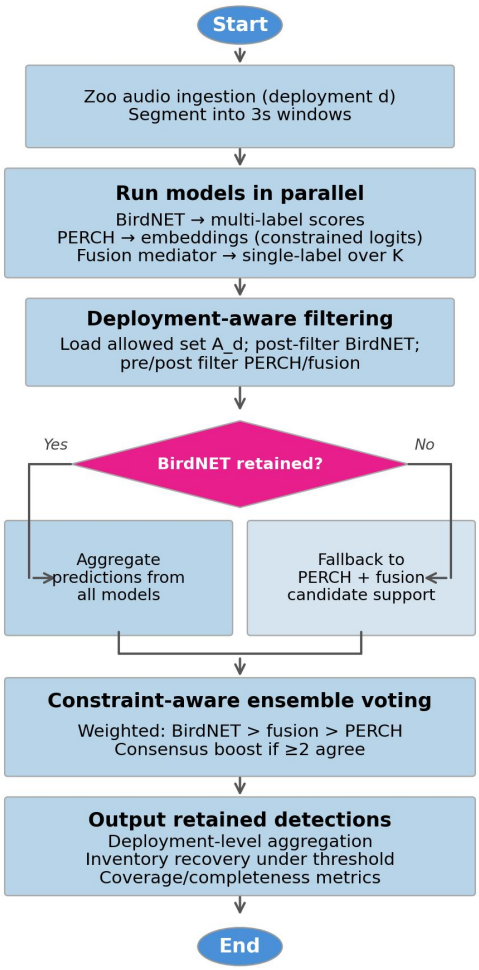



Fig. 2: Online inference pipeline (overview).

## VI. Segment-Level Agreement Results from Manual Annotation

Published guidance on BirdNET score interpretation establishes that confidence scores are unitless quantities—not probabilities—and that the relationship between score and detection accuracy is species-specific and context-dependent [15]. That guidance recommends two approaches for appropriate use: manual review of predictions down to a user-defined threshold, or conversion of scores to calibrated probabilities. We do not claim score calibration; instead, we adopted the first approach by designing an annotation protocol that samples detector outputs across the full confidence range and across all 16 zoo deployments. This stratified sampling enables direct evaluation of agreement patterns across confidence bins and quantification of empty prediction rates.

Three annotators, who are also the study authors, reviewed sampled detector outputs. The annotators have substantial familiarity with the dataset and model behavior developed through extended engagement with the corpus, but do not hold formal ornithological credentials. The resulting labels therefore constitute an operational ground truth for detector assessment in this study rather than an expert reference standard, which would ideally be provided by independent ornithologists. For each sampled segment, annotators assigned one of four labels: present (target species audibly confirmed), absent (target species judged not audible), uncertain (possible presence but insufficient acoustic evidence), or skip (segment not reliably assessable due to severe overlap, speech, noise, or other confounds). While the use of multiple annotators may reduce individual labeling bias, it does not guarantee freedom from systematic error [16].

The annotation effort produced 3,463 total segment-level labels. Of these, 2,541 annotations confirmed that the target species was audibly present. The remaining 922 annotations fall into two categories. A subset of 506 segments were labeled as absent, indicating that the annotator determined the target species was not present. The remainder were assigned skip or uncertain status, reflecting segments with ambiguous vocalizations, overlapping human speech, excessive background noise, or other conditions that precluded confident assessment. We exclude absent, skip, and uncertain annotations from agreement calculations to avoid conflating genuine model error with annotation ambiguity or confirmed absence.

An important consideration is species coverage. The default BirdNET model contains approximately 6,500 species, yet only 69 intersect with the zoos' inventories. Through targeted retraining, effective coverage is expanded to 87 species, thereby substantially increasing alignment with the annotated zoo taxa by including 18 species absent from default BirdNET, such as Bali Starling, Scarlet Ibis, Egyptian Vulture and Straw-necked Ibis, which are critical for zoo inventory monitoring. ARIA extends this label space to 89 species, with the annotation set containing two additional taxa (African Spoonbill and Red-crested Turaco) that are not represented in the retrained BirdNET vocabulary.

The quantitative comparison contrasts two configurations: default BirdNET constrained to species expected at each deployment, and the retrained BirdNET model with head-only fine-tuning on zoo-relevant audio. A key outcome is the rate of empty predictions, defined as segments where the model produces no species output after deployment-aware filtering and thresholding. For the constrained default model, 1,527 of the 2,541 confirmed-present segments yielded empty predictions, leaving 1,014 segments with retained output. Agreement denotes correspondence between the predicted species label and the annotator-assigned label. Under this criterion, the default model achieved 45.27% agreement. The retrained model reduced empty predictions to 470 segments, yielding 2,071 segments with retained output and achieving 79.58% agreement with annotator labels. This represents a 34.31 percentage point absolute improvement and approximately 76% relative improvement observed in this evaluation sample.

The improvement pattern is consistent across deployments: the retrained model yields higher agreement and fewer empty predictions at every zoo location evaluated. At the species level, 62 species in the annotated sample exhibit higher agreement

under the retrained model than under constrained default BirdNET. No species achieves 80% or higher agreement with the default model; under the retrained model, 36 species exceed this threshold, with 20 species achieving 100% agreement on their respective annotation samples. Conversely, 21 species show zero percent agreement under the default model despite being in its vocabulary, compared to only 3 species under the retrained model. These results demonstrate that head-only retraining constitutes a substantial first step toward reliable zoo soundscape monitoring. The magnitude of improvement—from effectively unusable detection rates to substantially higher agreement levels—indicates that making the classifier inventory-specific through targeted domain adaptation is essential for captive environment monitoring where default BirdNET systematically underperforms.

However, retraining does not resolve all failure modes. A few species with limited representation in training data, challenging acoustic signatures, or low vocal activity remain difficult to detect reliably. Therefore, species that vocalize infrequently or are rarely detected contribute fewer segments to the evaluation, and their performance estimates are correspondingly less stable. Examples include Lesser Kestrel and Mindanao Bleeding-heart Dove, both of which show negligible agreement even after retraining. These residual gaps indicate that head-only retraining, while effective for the majority of target taxa, does not eliminate coverage limitations for underrepresented or acoustically difficult species. Moreover, the persistence of 470 empty predictions on confirmed-present segments means that the retrained model still fails to produce any output for a meaningful fraction of true vocalizations—segments that would be entirely missed without complementary detection capability.

These empty predictions and the 79 % agreement rate motivate the ARIA architecture described in the preceding sections. The ARIA can recover predictions in segments where BirdNET produces no retained output or where confident but incorrect BirdNET predictions are adjusted through fusion-mediated correction, by leveraging PERCH v2 and learned score mediation. To characterize false-positive behavior, we evaluated ARIA predictions with definitive annotator labels: of 3,047 such segments, 2,541 contained annotator-confirmed target vocalizations while 506 were labeled as target-absent, yielding 83.4% precision (TP/(TP+FP)). This estimate is sample-based and derived from the same annotation corpus; generalization to the full distribution of acoustic conditions in continuous monitoring remains to be validated. Nevertheless, the pattern of results supports a two-stage interpretation: retraining provides a substantial step-change in detection coverage and agreement relative to constrained default BirdNET, while hybridization offers incremental improvement and extends detection capability to segments where the retrained model produces no output. The residual failure modes, concentrated in underrepresented species and acoustically challenging segments, indicate that further gains will require targeted data acquisition rather than architectural modification alone.

## VII. Full-Dataset Inventory Recovery and Detection Volume Analysis

The annotation study provides direct validation of detection agreement on a stratified sample. To complement this segment-level evaluation, we report detection volume and inventory recovery metrics across the complete unannotated corpus spanning 16 deployments and 999,177 segments. These metrics verify the consistent superiority of the retrained model over constrained default BirdNET and demonstrate the incremental coverage gains from hybridization. While these full-dataset metrics cannot assess prediction correctness without ground truth, they confirm that the improvements observed in the annotation sample generalize to operational deployment conditions.

### A. Default vs. Custom BirdNET Performance

At the operational threshold (confidence $\geq 0.5$), the retrained BirdNET model yields 577,539 detections versus 189,787 for constrained default (3.04$\times$ increase), with files containing any detection increasing from 170,526 to 453,692 (2.66$\times$ increase). Inventory coverage, defined as the fraction of expected cage taxa detected at least once per deployment, increases from a mean of 49.8% under default BirdNET to 89.6% under the personalized model. This improvement is consistent with the annotation study's finding that the retrained model expands effective species coverage from 69 to 87 taxa, including 18 species absent from default BirdNET's vocabulary. The personalized configuration achieves higher coverage in 15 of 16 deployments and matches the default in the remaining site. At the deployment completeness level, only 2 of 16 sites reach $\geq 80\%$ inventory recovery under default BirdNET, compared to 13 of 16 under the personalized model. Perfect inventory recovery (100%) increases from 1 to 7 deployments. These patterns corroborate the annotation-based findings: constrained default BirdNET leaves systematic inventory blind spots, whereas head-only retraining substantially closes these gaps across deployments.

### B. Custom BirdNET vs. ARIA Performance

ARIA extends coverage beyond the retrained BirdNET baseline. At the operational threshold, ARIA increases detections from 577,539 to 793,494 and files with any detection from 453,692 to 563,242. Species coverage expands from 87 to 89 taxa with the recovery of two previously undetected cage species (African Spoonbill and Red-crested Turaco). Mean inventory coverage rises modestly from 89.6% to 91.1%, with the gain concentrated in a single deployment achieving full recovery.

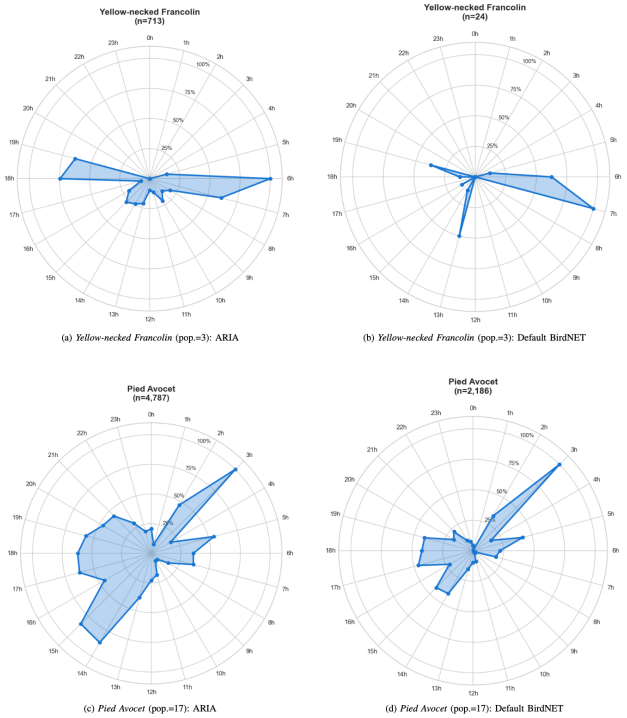
ARIA's advantage is particularly evident in BirdNET-null segments, defined as windows where the retrained BirdNET pipeline produces no retained prediction after deployment-aware filtering and thresholding. Across 139,195 such segments where both PERCH v2 and fusion model yield predictions, the two models agree on species identity in 52.5% of cases. When predictions differ, fusion systematically assigns lower confidence scores than PERCH v2, with stronger damping for cage taxa, consistent with learned mitigation of probability

inflation that occurs when PERCH v2's softmax distribution is constrained to deployment-specific label sets. These patterns indicate that the ARIA extends detection capability to segments where BirdNET evidence is absent, while applying learned reweighting that yields more conservative confidence profiles than PERCH v2 in ambiguous windows.

These behavioral patterns across the full corpus are consistent with the annotation study's findings, supporting the interpretation that retraining provides substantial gains over default BirdNET while hybridization extends coverage to segments where the retrained model produces no output.

## VIII. Activity-Profile Case Studies Across Population Sizes

To complement the corpus-level inventory and detection-volume analysis, Figs. 3 and 4 compare hour-of-day activity patterns between ARIA and the constrained default BirdNET baseline, using *retained* detections after the same deployment-aware species restriction, thresholding, and post-processing applied throughout the system. As illustrative case studies, we report one example deployment from each of four population-size regimes (3, 17, 52, and 153 individuals). The corresponding taxa for these deployments (Yellow-necked Francolin, Pied Avocet, Greater Flamingo, and Red-billed Quelea) are those for which the constrained default BirdNET baseline produced sufficiently frequent retained detections in the selected deployments to get interpretable hour-of-day profiles; in many other deployments or taxa, the default baseline either lacks the species in its label set or detections would be too sparse to support meaningful temporal summaries, making a direct activity-profile comparison uninformative.




(a) *Yellow-necked Francolin* (pop.=3): ARIA (b) *Yellow-necked Francolin* (pop.=3): Default BirdNET

(c) *Pied Avocet* (pop.=17): ARIA (d) *Pied Avocet* (pop.=17): Default BirdNET

Fig. 3: Activity profiles (pop.=3 and 17): ARIA vs. constrained default BirdNET.


(a) *Red-billed Quelea* (pop.=153): ARIA (b) *Red-billed Quelea* (pop.=153): Default BirdNET

(c) *Greater Flamingo* (pop.=52): ARIA (d) *Greater Flamingo* (pop.=52): Default BirdNET

Fig. 4: Activity profiles (pop.=52 and 153): ARIA vs. constrained default BirdNET.

Each radial subplot summarizes a single species at a single deployment: the angle corresponds to hour (0–23) and the radius shows the *relative* detection rate for that hour, normalized within the plot such that 100% denotes the peak hour for that species at that site; the title reports the total number of retained detections ($n$) used to form the hourly profile. Because normalization removes absolute scale, these plots emphasize temporal structure while $n$ conveys detection volume; importantly, these curves reflect model detection activity rather than ground-truth call rates or numbers of individuals, and they can be influenced by background noise and enclosure acoustics. Across the four population-size regimes—Yellow-necked Francolin (3), Pied Avocet (17), Greater Flamingo (52), and Red-billed Quelea (153)—ARIA yields higher retained detection volume and more stable (less spiky) diurnal profiles than default BirdNET. The largest relative gains occur in the smallest-population case (Francolin: $n$ increases from 24 to 713; $\sim$29.7$\times$), while larger populations still show substantial improvements (Quelea: 2,057 to 19,668; $\sim$9.6$\times$; Avocet: 2,186 to 4,787; $\sim$2.2$\times$; Flamingo: 26,649 to 41,750; $\sim$1.6$\times$), suggesting that the ARIA pipeline is particularly beneficial when baseline detections are sparse yet remains advantageous under higher-activity conditions. Notably, ARIA does not simply increase detections uniformly across the 24-hour cycle: the dominant peaks and troughs remain taxon-specific (e.g., pronounced morning peaks versus broader daytime activity), and the ARIA curves generally preserve the same primary activity regions as the default model while increasing detection volume and smoothing hour-to-hour variability. This qualitative consistency supports the

interpretation that the ARIA pipeline improves robustness without collapsing different taxa into a single, generic activity pattern, consistent with generalization across heterogeneous deployments and population-size regimes.

## IX. Conclusion

Zoo aviary soundscape monitoring constitutes an inventory recovery problem: confirming that known taxa vocalize under acoustically challenging conditions where general-purpose classifiers structurally underperform. ARIA addresses this through three integrated components: head-only retrained BirdNET for domain adaptation, PERCH v2 for broad taxonomic coverage and acoustic embeddings, and a learned fusion mediator that recovers detections in BirdNET-null segments while mitigating confidence artifacts induced by label-set constraints. Segment-level manual annotation demonstrates that domain adaptation alone improves agreement from below 50% to approximately 80% relative to constrained default BirdNET, with hybridization providing incremental gains on segments lacking BirdNET output. Full-corpus operational metrics across 16 deployments confirm these behavioral patterns at scale, showing substantial reductions in empty-prediction rates and improved inventory coverage. The consistency of these improvements across deployments with heterogeneous acoustic conditions and population sizes suggests that the ARIA architecture generalizes across operational zoo environments. Residual failure modes concentrate in underrepresented taxa and acoustically challenging segments, indicating that targeted data acquisition for these species represents the most promising direction for further improvement.

## Acknowledgments

We gratefully acknowledge receipt of media from the Cornell Lab of Ornithology | Macaulay Library. We are thankful to zoos for their collaboration and giving permission to collect audio data.

## References

[1] D. Stowell, "Computational bioacoustics with deep learning: a review and roadmap," *PeerJ*, vol. 10, p. e13152, Mar. 2022. [Online]. Available: http://dx.doi.org/10.7717/peerj.13152

[2] S. Kahl, C. M. Wood, M. Eibl, and H. Klinck, "Birdnet: A deep learning solution for avian diversity monitoring," *Ecological Informatics*, vol. 61, p. 101236, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1574954121000273

[3] F. E. Clark and J. C. Dunn, "From soundwave to soundscape: A guide to acoustic research in captive animal environments," *Frontiers in Veterinary Science*, vol. 9, p. 889117, 2022.

[4] B. van Merriënboer, V. Dumoulin, J. Hamer, L. Harrell, A. Burns, and T. Denton, "Perch 2.0: The bittern lesson for bioacoustics," *arXiv preprint arXiv:2508.04665*, 2025.

[5] Xeno-canto Foundation for Nature Sounds, "Xeno-canto – bird sounds from around the world," https://xeno-canto.org/, 2005.

[6] Cornell Lab of Ornithology, "Macaulay library," https://www.macaulaylibrary.org/, 2024.

[7] P. Lauha, P. Somervuo, P. Lehikoinen, L. Geres, T. Richter, S. Seibold, and O. Ovaskainen, "Domain-specific neural networks improve automated bird sound recognition already with small amount of local data," *Methods in Ecology and Evolution*, vol. 13, no. 12, pp. 2799–2810, 2022. [Online]. Available: https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.14003

[8] A.-C. Guei, S. Christin, N. Lecomte, and Hervet, "Ecogen: Bird sounds generation using deep learning," *Methods in Ecology and Evolution*, vol. 15, no. 1, pp. 69–79, 2024. [Online]. Available: https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.14239

[9] C. Pérez-Granados, "Birdnet: applications, performance, pitfalls and future opportunities," *Ibis*, vol. 165, no. 4, pp. 1068–1089, 2023.

[10] B. Ghani, V. J. Kalkman, B. Planqué, W.-P. Vellinga, L. Gill, and D. Stowell, "Impact of transfer learning methods and dataset characteristics on generalization in birdsong classification," *Scientific Reports*, vol. 15, p. 16273, 2025.

[11] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," pp. 2999–3007, 10 2017.

[12] H. Zhang, M. Cisse, Y. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," 10 2018.

[13] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proceedings of the 34th International Conference on Machine Learning*, ser. PMLR, vol. 70, Sydney, Australia, 2017, pp. 1321–1330.

[14] A. Márquez-Rodríguez, M. Ángel Mohedano-Munoz, M. J. Marín-Jiménez, E. Santamaría-García, G. Bastianelli, P. Jordano, and I. Mendoza, "A bird song detector for improving bird identification through deep learning: A case study from doñana," *Ecological Informatics*, vol. 90, p. 103254, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1574954125002638

[15] C. M. Wood and S. Kahl, "Guidelines for appropriate use of birdnet scores and other detector outputs," *Journal of Ornithology*, vol. 165, pp. 777–782, 2024.

[16] P. Nguyen Hong Duc, M. Torterotot, F. Samaran, P. R. White, O. Gérard, O. Adam, and D. Cazau, "Assessing inter-annotator agreement from collaborative annotation campaign in marine bioacoustics," *Ecological Informatics*, vol. 61, p. 101185, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1574954120301357