A TWO-STAGE CRNN REGRESSION FRAMEWORK FOR BIRD COUNTING

Technical Report

Jing Yang[1], *Siyi Li*[2], *Zihan Zheng*[1], *Heng Zhang*[1], *Youran Ni*[1], *Yuzhu Wang*[3], *Gongping Huang*[1]

[1] Wuhan University, Wuhan, Hubei 430072, China
[2] Beijing Jiaotong University, Beijing 100044, China
[3] Signal Processing Research Center, Tampere University, Tampere, Finland

ABSTRACT

Bird counting from passive acoustic recordings is challenging for chorus-prone species like the Greater flamingo due to synchronous group vocalizations. In this technical report, we describe our submitted system for the BioDCASE 2026 Bird Counting challenge. We propose a two-stage framework specifically tailored for the Greater flamingo. First, we construct a synthetic dataset and train a Mel-CRNN regressor to directly predict fragment-level call counts. Second, we aggregate fragment-level predictions into aviary-level statistical features and fit a linear regression model on the development set for final estimation. For other target species (Hadada ibis and Red-billed quelea), we directly adopt the official ARIA baseline strategies. The system pipeline, parameter configurations, and submitted file structures are detailed herein to ensure transparency and reproducibility.

***Index Terms*—** Bird population estimation, Passive acoustic monitoring, Greater flamingo, Count regression

1. INTRODUCTION

Passive acoustic monitoring (PAM) has become an essential tool for biodiversity research. However, the transition from species detection to individual counting remains methodologically challenging. For flock-calling species such as the Greater flamingo, when many individuals call within the same 3-second window, the acoustic scene becomes a continuous chorus rather than a sequence of distinguishable events. As a result, raw detection counts saturate, and the per-individual detection rate decreases as the population grows [1].

In the context of the BioDCASE 2026 Bird Counting challenge, the official baseline system [1] employs pre-trained detection models (e.g., BirdNET or ARIA [2]) combined with statistical feature extraction. The baseline results indicate varying performance across target species, with notably larger deviations observed for the Greater flamingo compared to Hadada ibis and Red-billed quelea. This disparity suggests that detection-based approaches may face limitations when applied to species with synchronous group vocalizations, motivating us to explore alternative estimation strategies.

To address this challenge, we propose a hybrid two-stage system. For the Greater flamingo, inspired by the direct estimate of size (DES) paradigm [3], we adopt a top-down regression strategy to mitigate the detection bottleneck. Rather than relying on a bottom-up approach that explicitly detects and counts individual vocalization events, our method directly predicts fragment-level call counts from acoustic features using a Mel-CRNN regressor trained on a synthetic dataset. These fragment-level estimates are subsequently aggregated into aviary-level statistical descriptors and mapped to final population counts via a calibrated linear model. For the remaining target species, we retain the official baseline strategies to ensure a comprehensive submission.

2. SYSTEM DESCRIPTION

2.1. External Data Usage and Compliance Statement

External Data: We utilized two types of external audio data: (1) Greater flamingo field recordings from Xeno-Canto [2] for extracting target call snippets; (2) Non-target species recordings from Xeno-Canto for generating interference sources. These external data were strictly used to construct the synthetic training dataset and were not part of the official challenge dataset [4].

Compliance Statement: We strictly declare that the evaluation set provided by the challenge organizers was never used for model training, hyperparameter tuning, architecture selection, or any form of validation during the system development phase. All model configurations and parameters were finalized solely based on the development set and synthetic data.

2.2. Synthetic Dataset Construction

Due to the lack of real-world recordings with precise call-count annotations, we construct a synthetic dataset by combining three audio components: isolated flamingo calls, background noise from development aviaries, and non-target interference calls.

Flamingo Call Snippets: We collected 429 field recordings of Greater flamingo from Xeno-Canto. Each recording was segmented into 3-second windows (50% overlap) and filtered using BirdNET and PERCH v2 (confidence ≥ 0.9). Isolated call snippets (0.12–1.50 s) were extracted via RMS energy envelope and adaptive boundary detection, yielding $2,587$ snippets. To prevent data leakage, splitting was performed by the original Xeno-Canto recording ID.

Background Audio: Background segments were derived from two sources: (1) Source A: Flamingo-free segments extracted from development aviary recordings [4].(2) Source B: Non-target interference calls extracted from Xeno-Canto recordings of other

[1] https://biodcase.github.io/challenge2026/task6

[2] https://xeno-canto.org

species, serving as additional acoustic interference during synthesis.

Data Splitting: To prevent data leakage, all source materials, including flamingo calls, background segments, and interference snippets, were split at the source level by their original recording IDs. This ensures that snippets derived from the same source recording never appear in both the training and validation sets.

Mixing Protocol: Prior to synthesis, all source audio components (flamingo calls, background segments, and interference calls) are uniformly resampled to 32 kHz to maintain consistency and reduce computational complexity. Consequently, each synthetic sample is generated as a mono 3-second clip at 32 kHz, formulated as:

$$x(t) = b(t) + \sum_{i=1}^{N} f_i(t) + \sum_{j=1}^{n_{\text{interf}}} u_j(t),$$

where $b(t)$ denotes background audio extracted from real development aviary recordings, $f_i(t)$ represents the i-th flamingo call, and $u_j(t)$ denotes non-target interference calls downloaded from Xeno-Canto. Here, N corresponds to the ground-truth label, representing the exact number of flamingo calls inserted into the clip. The number of interference calls n_{interf} is randomly sampled from $\{0, 1, 2, 3\}$. All audio components are mixed with independent random gains in the range $[-10, +10]$ dB. For the training set, we generate 150 synthetic samples for each count value ranging from 0 to 50. For the validation set, we generate 10 samples per count value.

2.3. Stage 1: Fragment-level CRNN Regression

Our CRNN regressor processes 3-second raw audio clips (32 kHz) through a sequential pipeline. First, the signal frontend computes a 128-band log-mel spectrogram (50–12,000 Hz, 1024-point FFT, 320-sample hop), applies amplitude-to-dB compression and per-sample z-score normalization, and utilizes SpecAugment during training, yielding a tensor of shape $(B, 128, T)$. The spectrogram is then fed into a CNN encoder consisting of three convolutional blocks. Each block contains two 3×3 convolutions, batch normalization, ReLU, 2×2 max-pooling, and dropout, progressively expanding the channel dimensions from 1 to 32, 64, and finally 128, while compressing the time axis. To condense spectral information while preserving temporal dynamics, we apply mean and max pooling along the frequency axis and concatenate the results, yielding a sequence of shape $(B, 256, T')$. This sequence is processed by a 2-layer bidirectional GRU (hidden size 128, dropout 0.25) to capture temporal dependencies, outputting 256-dimensional features per time step. Subsequently, temporal mean and max pooling are concatenated into a 512-dimensional global vector. Finally, this vector is passed through a regression head comprising LayerNorm, a linear projection ($512 \to 128$), ReLU, dropout (0.25), and a final linear layer ($128 \to 1$) terminated by a Softplus activation to ensure non-negative count predictions. The entire architecture contains approximately 1 M parameters.

Training Configuration: We employ the AdamW optimizer (lr=3×10^{-4}, weight decay=10^{-4}), batch size 32, and gradient clipping (max norm 5.0). The learning rate is reduced by a factor of 0.5 upon validation MSE plateau, with an early stopping patience of 10 epochs. We train the model using the square-root mean squared error ($\sqrt{\text{MSE}}$) loss function, defined as:

$$L_{\sqrt{\text{MSE}}} = \frac{1}{B} \sum_{i=1}^{B} (\sqrt{y_i} - \sqrt{\hat{y}_i})^2, \tag{1}$$

where y_i and $\hat{y}_i$ denote the true and predicted counts, respectively, and B is the batch size.

2.4. Stage 2: Aviary-level Aggregation and Estimation

For an aviary containing M 3-second fragments (resampled from the original 48 kHz to 32 kHz), the trained CRNN model infers each fragment t to obtain the predicted count $p_t \in [0, \infty)$. We compute the mean μ and standard deviation σ of all fragment predictions as the aviary-level statistical features.

Given that there are only 4 aviaries with non-zero Greater flamingo counts in the development set (dev_2, dev_4, dev_5, dev_6), we adopt a simple linear regression model to map these features to population counts:

$$\hat{C} = w_1 \cdot \mu + w_2 \cdot \sigma + b. \tag{2}$$

To maximize data utilization and ensure prediction stability, the model parameters (w_1, w_2, b) are fitted via ordinary least squares using all four development aviaries. For the evaluation set, the complete pipeline infers fragment-level counts using the trained CRNN model, aggregates them into (μ, σ), applies the linear model fitted on the development set (Eq. 2), and rounds the output to the nearest integer as the final population estimate.

2.5. Estimation for Other Target Species

For species other than Greater flamingo, our system directly adopts the official ARIA baseline strategies without any modifications. In Stage 1, species detection is performed using the official ARIA ensemble detection scheme (BirdNET, Fusion, and PERCH) with its default settings. For each recording, we select the top 3 species with the highest weighted voting scores without enforcing any species whitelist. In Stage 2, we directly apply the best-performing models from the official baseline priority list:

- **Hadada ibis**: We utilize the linear_coeff_bout_rate model, predicting population size based on the linear coefficient of the vocalization bout rate.
- **Red-billed quelea**: We employ the linear_coeff_cwr model, estimating abundance using the linear coefficient of the confidence-weighted detection rate (CWR).

3. RESULTS

Table 1 summarizes the system performance on the development set and the final predictions on the evaluation set. For the development set, the predicted counts and errors are obtained using a leave-one-out cross-validation (LOO-CV) setup. On the development set, our proposed CRNN-based framework for the Greater flamingo demonstrates strong performance, achieving a low mean absolute error (MAE) of 1.50 individuals with a maximum absolute deviation of only 2 across all four aviaries. This indicates that the direct regression approach effectively captures the underlying vocal density patterns even in the presence of overlapping calls. Across all target species on the development set ($N = 8$), the system achieves an overall MAE of 0.75, RMSE of 1.12, R^2 of 0.9996, and MAPE of 1.0%, demonstrating high prediction accuracy and strong correlation with true population sizes. For the evaluation set, the final linear regression model is fitted on all available development aviaries, and the ground-truth values are not available.

Table 1: Performance on the development set and predictions on the evaluation set. For the development set, the error (Err) represents the difference between the predicted and true counts. For the evaluation set, ground-truth values are not available, hence True and Err are marked as "–".

Species	Best Model	Aviary	True	Pred	Err
Greater flamingo	CRNN	dev_aviary_2	107	108	+1
Greater flamingo	CRNN	dev_aviary_4	161	159	-2
Greater flamingo	CRNN	dev_aviary_5	52	54	+2
Greater flamingo	CRNN	dev_aviary_6	52	51	-1
Hadada ibis	linear_coeff_bout_rate	dev_aviary_2	6	6	+0
Hadada ibis	linear_coeff_bout_rate	dev_aviary_4	4	4	+0
Red-billed quelea	linear_coeff_cwr	dev_aviary_1	153	153	+0
Red-billed quelea	linear_coeff_cwr	dev_aviary_3	61	61	+0
Red-billed quelea	linear_coeff_cwr	eval_aviary_1	–	149	–
Red-billed quelea	linear_coeff_cwr	eval_aviary_2	–	56	–
Greater flamingo	CRNN	eval_aviary_4	–	40	–
Hadada ibis	linear_coeff_bout_rate	eval_aviary_6	–	1	–
Hadada ibis	linear_coeff_bout_rate	eval_aviary_7	–	1	–
Greater flamingo	CRNN	eval_aviary_8	–	184	–

4. CONCLUSION

This technical report presents a detailed description of our submitted system for the BioDCASE 2026 Bird Counting challenge. To address the challenge of raw detection counts saturating when individuals vocalize synchronously, we designed a hybrid two-stage pipeline that integrates a top-down regression-based CRNN framework for the Greater flamingo with the official baseline strategies for other target species. By mitigating the detection bottleneck and leveraging fragment-level call count regression, our system is specifically designed to handle the acoustic complexities inherent to overlapping vocalizations. Furthermore, this document details the synthetic data construction, model architectures, hyperparameter configurations, and submission formats, ensuring transparency and reproducibility to support future research and community benchmarking.

5. SUBMITTED SYSTEM AND OUTPUT FORMAT

This section details the submitted system files and their corresponding metadata to ensure clear mapping between the technical report and the submitted results.

Submission Package: The submission is provided as a compressed archive named Yang_task6_submission, containing the following components:

- Yang_task6.technical_report.pdf: This technical report describing the system methodology.
- predictions_task6_Yang.csv: The CSV file containing population estimates for all evaluation recordings.
- Yang_task6_1.meta.yaml: The metadata file describing system configuration and submission details.
- code/: A directory containing the complete inference pipeline, including all necessary scripts, pre-trained model weights, and a comprehensive README file with detailed instructions for reproduction.

Output File Structure: The population estimates are provided in predictions_task6_Yang.csv with the following format:

- aviary_id: The unique identifier for each evaluation aviary (e.g., eval_aviary_1).
- species: The common name of the target species (Greater flamingo, Hadada ibis, or Red-billed quelea).
- predicted_count: The estimated integer population count.

6. REFERENCES

[1] E. Argın, A. Härmä, and A. Arslan-Dogan, "BioDCASE 2026 Bird Counting Baseline: Avian Population Estimation from Passive Acoustic Recordings," 2026. [Online]. Available: https://github.com/ml4biodiversity/biodcase-population-estimation

[2] E. Argın, B. Amado Pereira da Costa, A. Härmä, and A. Arslan-Dogan, "ARIA: Acoustic Recognition for Inventory in Aviaries," in *Proceedings of the IEEE World Congress on Computational Intelligence (WCCI) / International Joint Conference on Neural Networks (IJCNN)*, 2026, accepted, to appear.

[3] A. Arslan-Dogan and A. Härmä, "Counting without seeing: Toward acoustic population estimation from unsupervised audio features," *Proceedings of the 19th International Joint Conference on Biomedical Engineering Systems and Technologies*, 2026. [Online]. Available: https://api.semanticscholar.org/CorpusID:286406452

[4] E. Argın, A. Härmä, and A. Arslan-Dogan, "BioDCASE 2026 Bird Counting: Avian Population Estimation from Passive Acoustic Recordings," 2026. [Online]. Available: https://huggingface.co/datasets/Emreargin/BioDCASE2026_Bird_Counting