SG-MSM: SNR-Guided Data Augmentation for Mitigating Domain Shift in Avian Bioacoustics

Anonymous authors
Paper under double-blind review

Abstract

Passive acoustic monitoring (PAM) for bird classification suffers from a severe domain shift between focal training data and passive deployment environments, a challenge exacerbated by the prohibitive cost of annotating target data. Extensive benchmarking has shown that standard architectures and data augmentations struggle to generalize across this focal-to-passive gap. We propose SNR-Guided Multi-Signal Mix (SG-MSM), a multi-stage data augmentation pipeline that explicitly mitigates the focal-to-passive domain shift. We evaluate SG-MSM across six diverse PAM environments using an EfficientNet-B3, achieving a mean ROC-AUC of **0.884**. This outperforms the comprehensive augmentation pipeline established by the BirdSet benchmark and the Perch 1.0 foundation model (860K recordings) across all domains, despite training on orders of magnitude less data (3.6K–28K samples). SG-MSM proves highly competitive with the state-of-the-art Perch 2.0 and BirdMAE models. Component-wise ablation studies isolate the impact of each synthesis stage, demonstrating relative gains of **+17.7%** in ROC-AUC, **+68.5%** in cmAP, and **+69.1%** in Top-1 Accuracy over focal-only training. Source code will be made publicly available upon acceptance.

1 Introduction

Audio classification refers to the task of assigning audio signals to predefined categories based on their acoustic properties (Hershey et al., 2017). In recent years, this technique has gained prominence in biodiversity monitoring, where it plays a critical role in the automated detection and tracking of wildlife species (Ramesh et al., 2023). The integration of audio classification into ecological workflows enables more efficient and accurate monitoring, offering near-real-time insights into ecosystem health. For example, automated systems can process large volumes of acoustic data collected by PAM devices deployed in remote environments, facilitating long-term biodiversity assessment. By automating species identification based on vocalizations, these systems support researchers in making evidence-based decisions to mitigate threats such as habitat degradation and climate change. This application of acoustic AI contributes meaningfully to global sustainability objectives, particularly in the preservation of biodiversity (Chowfin & Leslie, 2021).

Existing models like Perch (Hamer et al., 2023) for bird audio classification are trained on focal recordings[1] and do not work as expected on PAM data (Denton et al., 2022). Figure 1 and Table 1 illustrate this limitation: while Perch models miss the background species (whwbec1) despite confidently detecting the foreground (cintin1), SG-MSM achieves balanced detection of both, effectively mitigating SNR bias.

One of the core challenges in PAM lies in the *domain shift* between labeled focal recordings (source domain) and unlabeled passive recordings (target domain). Focal recordings contain consistently high SNR bird vocalizations across all species, while passive recordings exhibit varied and predominantly low SNR levels with target signals often masked by environmental noise. A seemingly straightforward solution would be to simply collect and annotate passive recordings from the target deployment site. However, recent literature confirms this is practically prohibitive: annotating passive data requires "hundreds of hours of expert annotation

[1]These recordings use a directional microphone towards the source of the acoustic event, capturing sequences of primary and occasionally secondary calls.

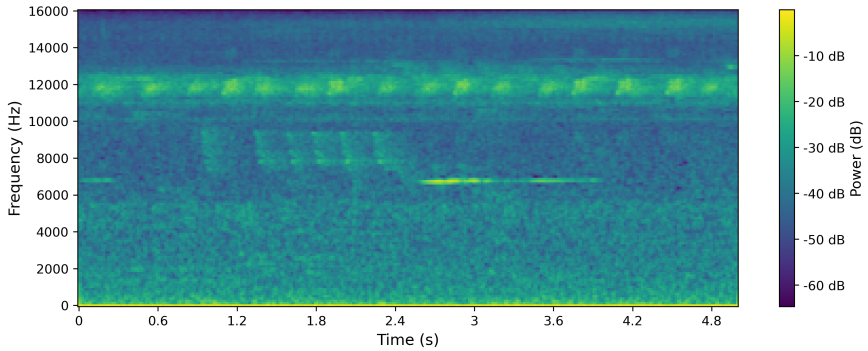

Figure 1: Mel-spectrogram of passive sample from PER domain: background *whwbec1* appears at 1.0–2.4s, followed by foreground *cintin1* at 2.5–4.0s.

Table 1: Multilabel classification on passive sample from PER domain (Figure 1). Perch probabilities filtered to PER regional domain. Ground truth species marked with (GT).

SG-MSM (Ours)		**Perch 2.0**		**Perch 1.0**	
Species	**Prob.**	**Species**	**Prob.**	**Species**	**Prob.**
cintin1 (GT)	**99.4%**	cintin1 (GT)	56.2%	cintin1 (GT)	93.3%
whwbec1 (GT)	**98.3%**	bartin2	4.8%	bartin2	17.0%
fepowl	39.2%	whwbec1 (GT)	0.9%	littin1	3.8%
bartin2	30.2%	littin1	0.6%	grfdov1	1.3%
greant1	5.0%	grasal3	0.5%	whwbec1 (GT)	0.4%

effort" (Hamer et al., 2023), local PAM subsets cause drastic label shifts that render them unsuitable for large-scale training (Rauch et al., 2025b), and the resulting annotated sets are generally too small for robust generalization (Van Merriënboer et al., 2024). Consequently, models must rely on focal recordings, which exposes a fundamental learning bias: models learn to associate high SNR with positive labels. This causes models to systematically underperform on lower-SNR background birds during cross-domain deployment in real-world PAM scenarios.

To bridge this focal-to-passive domain gap, standard data augmentations (e.g., Mixup, Time/Frequency Masking) have been extensively explored. However, recent benchmarks reveal that models that use these standard augmentations still face "substantial difficulties" in real-world PAM, stressing the "need for ongoing research with domain-specific approaches to address the complexities of model robustness" under covariate shift (Rauch et al., 2025b). Conversely, massive supervised pretraining (e.g., Perch 2.0 (van Merriënboer et al., 2025) with 1.5M recordings) achieves strong generalization, yet the substantial resource requirements are often inaccessible for many conservation scenarios.

Our framework, SNR-Guided Multi-Signal Mix (SG-MSM), explicitly bridges the focal-to-passive gap by employing a novel SNR-guided mixing algorithm (SG-Mix) to blend focal calls with mined passive noise. By utilizing unlabeled noise directly from the target deployment site, our approach inherently accounts for both the SNR distribution shift and the specific environmental noise profile of the target domain. SG-MSM generates passive-like multilabel training data that mitigates the SNR-confidence bias while preserving discrete multi-label definitions. This positions SG-MSM as a model-agnostic data augmentation framework that explicitly answers the call for domain-specific robustness under low-resource constraints.

2 Related Work

Depending on the recording type (*focal or PAM, also known as soundscape recording*), the classification problem is *multi-class* or *multi-label*, respectively. Focal recordings are typically formulated as multi-class

tasks (one dominant bird species per recording). In the latter case, PAM recordings require multi-label classification because they may capture a complex soundscape of primary events alongside distant, overlapping background species. While PAM is critical for scalable biodiversity monitoring, automated classification in this domain remains significantly more challenging than in focal recordings due to labeled data scarcity, high annotation costs, and complex soundscapes (Rauch et al., 2025b).

Large-Scale Pretraining Given the scarcity of labeled bioacoustic data, transfer learning has become a standard paradigm. Supervised pretraining on large-scale avian datasets (e.g., BirdNET (Kahl et al., 2021), Perch (Hamer et al., 2023)) or fine-tuning speech models like Wav2Vec 2.0 (Swaminathan et al., 2024) has proven effective for generating robust embeddings. In the domain of large-scale pretraining, Perch 1.0 (Ghani et al., 2023) utilized EfficientNet on weakly-labeled avian recordings to learn robust embeddings. Recently, Perch 2.0 (van Merriënboer et al., 2025) expanded this capability by training on massive multi-taxa datasets. However, models trained on general audio domains, such as AudioSet, often exhibit poor transferability to fine-grained bioacoustic tasks (Ghani et al., 2023; Nolan et al., 2023). Notably, Rauch et al. Rauch et al. (2025a) demonstrate that even general-purpose self-supervised learning (SSL) models struggle to capture the fine-grained discriminability required for bird classification, necessitating domain-specific pipeline adaptations such as Bird-MAE. Recent BirdCLEF competitions (Kahl et al., 2024; Cañas et al., 2025) established iterative pseudo-labeling as a dominant strategy, but such self-training methods risk amplifying errors when the domain shift is severe (Boudiaf et al., 2023; Sun et al., 2025).

Domain Adaptation and Data Augmentation Various domain adaptation and data augmentation methods have been explored to address these severe domain shifts and data limitations. Unsupervised separation techniques like MixIT (Wisdom et al., 2020) disentangle overlapping vocalizations to improve classification (Denton et al., 2022), though aggressive separation risks suppressing contextual cues or fragmenting calls. Active learning offers another route to reduce annotation burdens while maintaining robustness in dynamic environments (Bellafkir et al., 2023; Rauch et al., 2024). At the architectural level, Heinrich et al. Heinrich et al. (2025) proposed AudioProtoPNet, an interpretable prototype-based network. They demonstrated that prototype classification heads with a ConvNeXt backbone can outperform standard linear heads (e.g., Perch 1.0) on the BirdSet benchmark in both mAP and ROC-AUC, while offering per-species insights. Benchmarks like BirdSet (Rauch et al., 2025b) show that standard augmentations (e.g., Mixup (Zhang et al., 2017), Frequency and Time Masking (Park et al., 2019)) combined with large-scale pretraining still face "substantial difficulties" in effectively bridging the focal-to-passive gap (see Table 3 and §C.4 in Rauch et al. (2025b)).

Addressing the Gap Indeed, Perch 2.0 posits that finding the right data augmentations for bioacoustics remains an open research question (van Merriënboer et al., 2025). In contrast, under low resource constraints, SG-MSM directly answers this question through a *targeted data augmentation pipeline*. By employing a novel SNR-guided mixing algorithm (SG-Mix) to blend focal calls with mined passive noise, SG-MSM preserves discrete multi-label definitions to accurately model the complex acoustic foregrounds and backgrounds of real-world PAM (detailed in §3.4). While Perch 2.0 validates the "Bitter Lesson" (Sutton, 2019)—confirming that massive supervised pretraining yields greater generalization (van Merriënboer et al., 2025)—we investigate whether highly effective data augmentation can offer a viable alternative when such scale is inaccessible. Our work challenges the dependency on brute-force scaling by exploring a model-agnostic, SNR-guided data augmentation framework that enables robust classification without the computational prerequisites of foundation model training.

3 Methodology

3.1 Problem Formulation

A major challenge in classifying birds in a PAM environment arises from the SNR distribution shift: focal recordings exhibit concentrated high-energy bird calls, while passive recordings contain the same species at significantly lower SNR levels with complex background interference patterns, compounded by the scarcity of labeled focal data (i.e., a low-resource setting) for any given target environment.

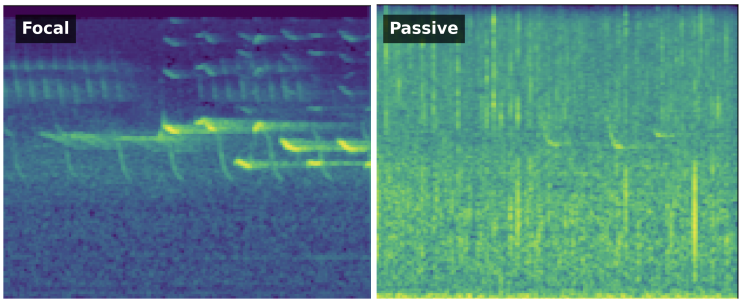

Figure 2: Comparison of Balori bird vocalizations spectrogram.

As seen in Figure 2, the vocalization of the same bird differs significantly between focal and PAM environments, causing models trained on focal data to fail on PAM data (Rauch et al., 2025b).

Concretely, we operate in a setting defined by a significant data asymmetry: given a low-resource set of labeled focal recordings $\mathcal{F} = \{(x_i^f, y_i)\}_{i=1}^{N_f}$ where x_i^f represents high-SNR audio segments and y_i denotes the primary species label, and an abundant set of unlabeled passive recordings $\mathcal{P} = \{x_j^p\}_{j=1}^{N_p}$ from target acoustic environments, where N_f and N_p denote the number of focal and passive recordings, and $N_f \ll N_p$. Our objective is to leverage the background noise and SNR distributions of the abundant, unlabeled data ($\mathcal{P}$) to learn a robust classifier f_θ that effectively generalizes from the scarce, labeled data ($\mathcal{F}$) to the target domain without requiring expensive passive data annotation. This setup simulates a site-specific calibration scenario, where a PAM system is adapted to the specific noise profile of a deployment site using available unlabeled audio, prior to the active detection phase.

3.2 Focal Recording Preprocessing

We utilize recordings from the BirdSet train set (Rauch et al., 2025b). To localize potential bird vocalizations within these weakly labeled files, we implement an energy-based event detection pipeline. We process these weakly labeled files using a step-by-step logic: (1) we apply a 300 Hz high-pass filter and denoising solely to compute a clean frame-level energy profile for event detection; (2) we isolate local energy peaks using `scipy.signal.find_peaks`; (3) we process these peaks in descending order of magnitude; and (4) we extract a centered 5-second window for each peak directly from the original, unfiltered audio, retaining candidate windows only if they share less than 50% temporal overlap with previously selected segments. This iterative selection terminates when peak energy falls below the recording's mean energy level. During training, we randomly sample one segment from this pool of extracted events, leveraging the high probability that dominant acoustic signals in focal recordings correspond to the labeled target species.

3.3 Background Noise Mining and SNR Categorization

We mine background noise from continuous raw passive BirdSet recordings by segmenting files into 1-minute windows and extracting the lowest-SNR 5-second segment from each. By isolating the minimum-energy segment within each extended temporal window, we leverage the natural sparsity of avian vocalizations to probabilistically capture the ambient environmental background.

For each 1-minute window, we compute SNR values using overlapping 5-second analysis windows with 2.5-second steps. The computed SNR values are partitioned into three 5dB ranges above the mined background SNR. We then average the SNR values within each range to obtain three representative values per extracted segment. To establish data-driven SNR categories, we apply K-means clustering ($K = 3$) to all representative values across recordings to generate *HIGH* (S_1), *MEDIUM* (S_2), and *LOW* (S_3) SNR categories. For the extracted noise segments exhibiting heavy environmental masking (empirically bounded by spectral entropy < 0.35), we introduce a fourth category (S_4) to model *VERY LOW* SNR conditions, defined as $S_4 = \text{background_SNR} + \mathcal{U}(0.1, 0.3)$ dB to simulate near-noise-floor signals. We choose K-means clustering over a grid search because it provides a data-driven approach to establishing SNR categories. K-means

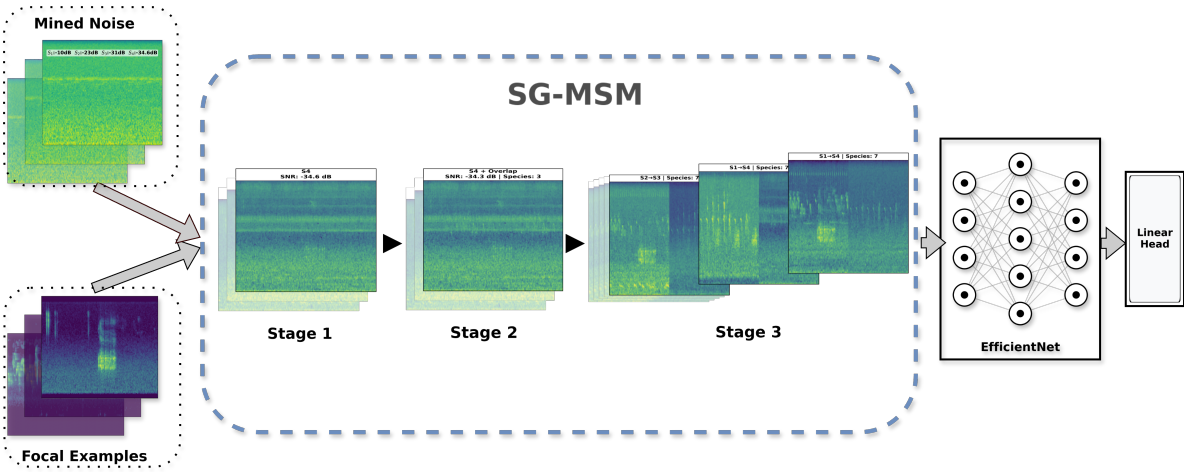

Figure 3: The proposed SG-MSM framework. **Left:** Input data sources (Mined Noise and Focal Examples) are combined. **Center:** The three-stage mixing process (SG-MSM) creates the training batch via Stage 1 (Multi-SNR Domain NoiseMix), Stage 2 (Iterative Multi-Label Overlapping), and Stage 3 (Cross-SNR Spectrogram Splicing). **Right:** The processed batch is fed into EfficientNet for classification.

automatically adapts the thresholds directly to the target domain's unique noise distribution. This ensures these data-driven categories accurately capture the specific environmental profile pattern of the deployment site.

3.4 SG-MSM Framework

As illustrated in Figure 3, our framework operates through three sequential stages: (1) Multi-SNR Domain NoiseMix, (2) Iterative Multi-Label Overlapping, and (3) Cross-SNR Spectrogram Splicing. Before detailing these stages, we introduce the core SG-Mix algorithm that enables the precise signal mixing required for our synthesis pipeline.

3.4.1 SNR-Guided Mixing Algorithm (SG-Mix)

The SNR-Guided Mixing Algorithm enables precise SNR control when mixing focal recordings with background noise to achieve target acoustic characteristics that match passive recording conditions. Given a focal recording, background noise, and target SNR level, this algorithm determines the optimal mixing ratio α that synthesizes passive-like recordings while preserving species label information. This synthesis process is mathematically grounded in the Signal-to-Noise Ratio (SNR), formally defined as:

$$SNR_{dB} = 10 \log_{10} \left(\frac{P_{signal}}{P_{noise}} \right) \tag{1}$$

Algorithm 1 employs iterative SNR calculation with binary search convergence to determine the mixing ratio. The approach utilizes (1) stationary noise reduction for signal-to-noise energy estimation via `noisereduce`, a tool empirically validated as fast and robust for general-domain time-series signals, including bioacoustics (Sainburg et al., 2020; Sainburg & Zorea, 2025), and (2) binary search operating in $\mathcal{O}(\log n)$ time to efficiently locate the mixing parameter α that satisfies the target SNR. To further optimize computational throughput, our accelerated implementation (Algorithm 3 in Appendix A) utilizes component caching, requiring exactly two `noisereduce` operations per mix. The core logic is outlined in Algorithm 1.

Algorithm 1 SG-Mix

Input: Focal x_f, Background b, Target SNR μ, Tolerance ϵ
Output: Mixed audio x_{mix}

Function GetSNR(y):
  $y_{signal} \leftarrow$ noisereduce(y, stationary=True)
  $y_{noise} \leftarrow y - y_{signal}$
  return $10\log_{10}(\sum y_{signal}^2 / \sum y_{noise}^2)$

Function Mix(α):
  $y \leftarrow \alpha \cdot x_f + (1-\alpha) \cdot b$
  return $y / \max(|y|)$ {Normalize & Mix}

Optimization (Binary Search):
Find optimal mixing ratio α^* via binary search such that:
  $|\text{GetSNR}(\text{Mix}(\alpha^*)) - \mu| \leq \epsilon$

return Mix(α^*)

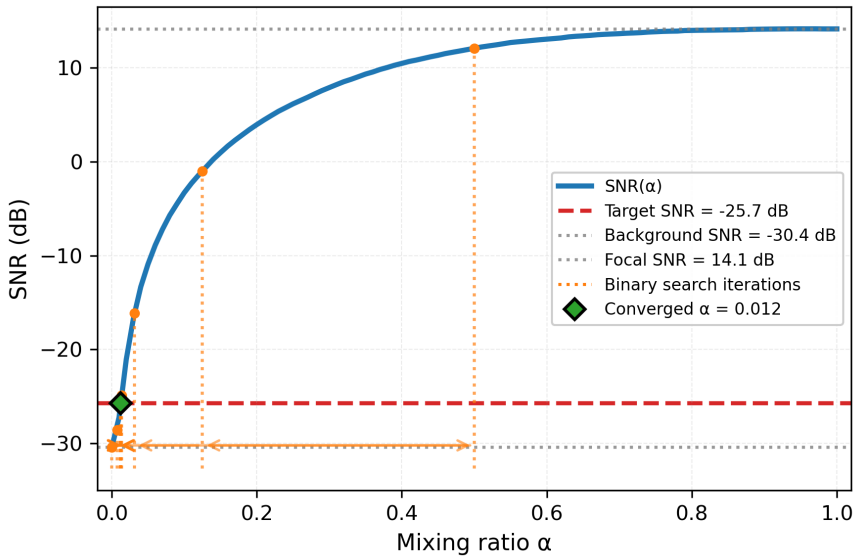

Figure 4: Binary search convergence to target SNR. Mathematically a bisection method: the Intermediate Value Theorem guarantees mixing ratio α exists (target SNR between SNR bounds), and bisection converges to it.

3.4.2 Stage 1: Multi-SNR Domain NoiseMix

This stage synthesizes passive-like recordings across varying SNR categories by mixing focal recordings with mined background noise (as described in Section 3.3).

For each focal recording x_i^f, we generate $|c|$ SNR variants by mixing with category-appropriate background noise (visualized in Figure 3):

$$x_i^{(c)} = \text{SG-Mix}(x_i^f, b_c, S_c(W_j)) \tag{2}$$

where $c \in \{S_1, S_2, S_3, S_4\}$, b_c is randomly selected background noise from category c that has lower SNR than the focal recording x_i^f, and $S_c(W_j)$ is the target SNR from the selected background recording W_j. Stage

Algorithm 2 Iterative Multi-Label Overlapping

Input: Base Mix x_{curr}, Labels y_{curr}, Relative Margin ρ
Output: Multi-species Mix x_{final}, Labels y_{final}

1: Determine $N_{species}$ (random $2 \ldots N_{max}$)
2: **for** $k = 1$ to $N_{species} - 1$ **do**
3: $\text{SNR}_{curr} \leftarrow \text{GetSNR}(x_{curr})$
4: Select secondary sample (x_{sec}, y_{sec}) such that:
5: $\text{SNR}(x_{sec}) > \text{SNR}_{curr}$ {Ensure distinctness}
6: $x_{curr} \leftarrow \text{SG-Mix}(x_{sec}, x_{curr}, \text{target} = \text{SNR}_{curr} + |\rho \cdot \text{SNR}_{curr}|)$ {Ensure absolute positive margin applied}
7: $y_{curr} \leftarrow y_{curr} \cup y_{sec}$ {Union labels}
8: **end for**
9: **return** x_{curr}, y_{curr}

1 expands each source focal recording into $|c|$ independent training examples (one per SNR category), where $|c|$ denotes the number of categories used. Stages 2–3 subsequently operate on these expanded examples without further size increase.

3.4.3 Stage 2: Iterative Multi-Label Overlapping

This stage simulates co-occurrence by iteratively mixing secondary signals. The core mixing logic is summarized in Algorithm 2.

This stage simulates passive recording conditions where multiple bird species co-occur with weaker signal strengths, training models to detect both dominant foreground and successively fainter background species. Since both the current mixture and secondary examples are peak-centered from preprocessing, this iterative mixing creates natural temporal overlap of vocalizations at the 2.5-second center point. The iterative mixing strategy preserves label integrity: simultaneous mixing with randomized weights can disproportionately amplify higher-SNR vocalizations while suppressing lower-SNR ones, rendering background species imperceptible despite valid annotations. Sequential addition with a small dynamic relative margin ($\rho = 1\%$) ensures each secondary species maintains acoustic presence, preserving label validity and detectability while mimicking foreground-background structure in PAM.

3.4.4 Stage 3: Cross-SNR Spectrogram Splicing

This stage performs mel-spectrogram-level splicing operations across different SNR categories. Following Stage 2, batched multi-SNR audio variants are converted to colorized mel-spectrograms using the feature engineering approach from Rasendiran R & Maurya (2025). To keep this framework self-contained, we note that a spectrogram is a visual representation of the frequency spectrum over time, mathematically defined as the squared magnitude of the Short-Time Fourier Transform (STFT) of the discrete time-domain signal $x[n]$:

$$S(\tau, k) = \left| \sum_{n=0}^{N-1} x[n + \tau H] w[n] e^{-j2\pi kn/N} \right|^2 \tag{3}$$

where τ is the time-frame index, k is the frequency bin index, H is the hop length, and $w[n]$ is a windowing function of length N. A mel-spectrogram subsequently maps these linear frequency bins f to the mel scale m using triangular filterbanks based on the transformation:

$$m = 2595 \log_{10} \left(1 + \frac{f}{700} \right) \tag{4}$$

For each batch, with a probability p_{csc}, the set of SNR categories is randomly shuffled into a sequence $t = (S_1, \ldots, S_4)$. We generate mixed samples by splicing each instance i in the current SNR category t_k

(where k denotes the kth category) with a deranged instance d from the next category t_{k+1} (where indices wrap via modulo $|t|$):

$$x_i^{t_k \to t_{k+1}} = \text{Splice}(x_i^{(t_k)}, x_{d_{k,i}}^{(t_{k+1})}) \tag{5}$$

where $d_{k,i}$ represents the independent random derangement index for the transition, ensuring that species labels are unified as $y_i^{\text{mix}} = y_i \cup y_{d_{k,i}}$. The Splicing operation constructs a composite 5-second mixed spectrogram by joining the center 2.5-second temporal segment of the current SNR variant with the corresponding segment of the next SNR category variant. By artificially exposing the model to disparate SNR features within a single training instance, splicing forces the network to detect background and foreground species without SNR bias. This mechanism directly mimics the acoustic complexity of real-world PAM (e.g., the co-occurrence of faint background and loud foreground species shown in Figure 1), effectively preventing the suppression of secondary species detection.

3.5 Network Architecture and Training

We employ EfficientNet-B3 as the backbone architecture with a linear classification head for multilabel classification using binary cross-entropy loss. Unlike traditional Convolutional Neural Networks (e.g., ResNet) that scale network dimensions arbitrarily, EfficientNet utilizes a principled compound scaling method that uniformly scales depth, width, and resolution. Combined with mobile inverted bottleneck convolutions (MBConv), this architecture provides a good balance of representational capacity and parameter efficiency Tan & Le (2019).

We select this lightweight configuration to meet edge PAM constraints. Furthermore, EfficientNet-B3 is a proven, standard backbone in bioacoustics that specifically aligns with the Perch 2.0 encoder baseline. This controlled setup ensures that our observed performance gains stem entirely from the SG-MSM data augmentation pipeline rather than architectural advantages. Training utilizes the complete SG-MSM synthesized dataset from Stages 1-3 (visualized in Figure 3).

3.6 Theoretical Foundations

Our SG-MSM framework mitigates the focal-to-passive domain shift under low-resource constraints (3.6K–28K labeled examples per environment). We ground our approach in domain generalization theory and data augmentation principles.

Domain Alignment via Targeted Augmentation Let $\mathcal{D}_S$ denote the source (focal) domain and $\mathcal{D}_T$ denote the target (passive) domain. Following Ben-David et al. (2010), target domain error is bounded by:

$$\epsilon_T(h) \leq \epsilon_S(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T) + \lambda \tag{6}$$

where $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T)$ is the $\mathcal{H}\Delta\mathcal{H}$-divergence, which measures the distance between the source and target distributions with respect to the hypothesis space $\mathcal{H}$, and $\lambda = \min_{h \in \mathcal{H}}\{\epsilon_S(h) + \epsilon_T(h)\}$ quantifies task adaptability. While domain shift in PAM arises from multiple interacting factors (Rauch et al., 2025b), we identify SNR degradation—driven by sound source distance and location-specific background noise (illustrated in Figure 5)—as a dominant distributional shift. Consequently, we posit that λ remains small since focal and passive recordings contain identical bird species. Our framework reduces the $\mathcal{H}\Delta\mathcal{H}$-divergence term strictly in data space prior to training. By constructing an augmented source distribution $\hat{\mathcal{D}}_S$, SG-MSM explicitly aligns the training data with the target passive domain's SNR distribution and co-occurrence patterns.

Augmentation-Driven Generalization Consistent with Hoffer et al. (2020), augmented instances of the same sample produce correlated gradients, resulting in controlled variance reduction relative to independent samples. Our framework generates training examples where each focal recording is represented across

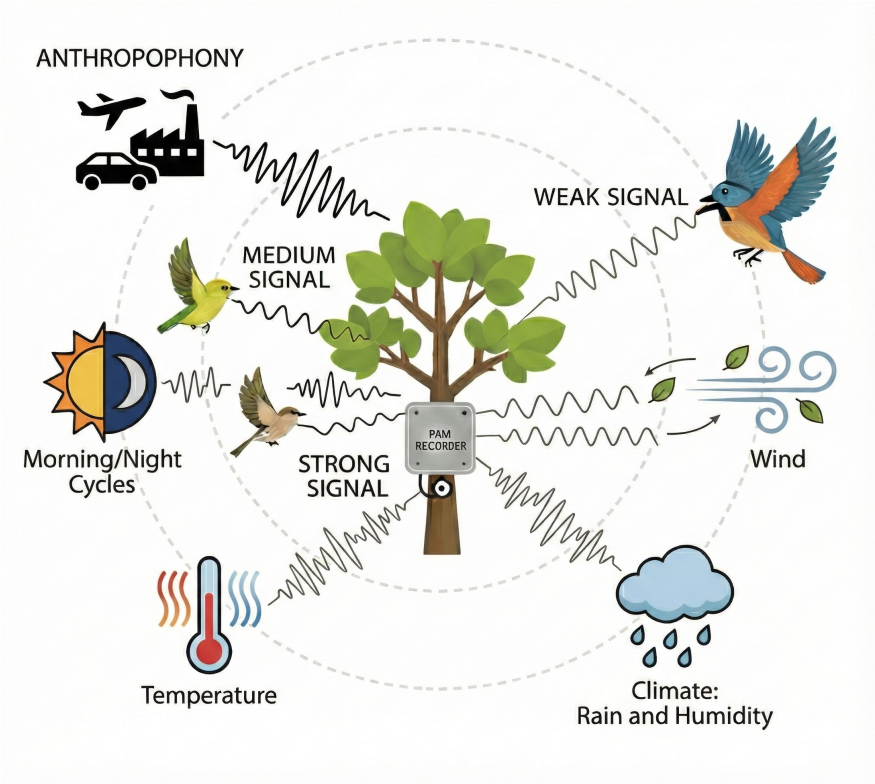

Figure 5: PAM scenario illustration. Bird signal strength attenuates over distance, while environmental factors and anthropophony introduce variable background noise. This results in diverse SNR levels and overlapping vocalizations from multiple species, constituting the multilabel challenge addressed by SG-MSM.

multiple SNR variants $\{x_i^{(S_1)}, x_i^{(S_2)}, x_i^{(S_3)}, x_i^{(S_4)}\}$, satisfying:

$$\mathbb{E}[\mathrm{Corr}(\nabla_\theta \ell(\theta, x_i^{(S_j)}, y_i), \nabla_\theta \ell(\theta, x_i^{(S_k)}, y_i))] > \mathbb{E}[\mathrm{Corr}(\nabla_\theta \ell(\theta, x_i, y_i), \nabla_\theta \ell(\theta, x_m, y_m))] \tag{7}$$

for $j \neq k$ and $i \neq m$. Here, θ represents the model parameters, ℓ is the objective loss function, and $\nabla_\theta \ell$ denotes the computed gradients. This inequality formally states that the expected gradient correlation between two differently augmented SNR variants of the same focal recording (S_j, S_k) is strictly greater than the correlation between two entirely different training samples (x_i, x_m). This gradient correlation property, arising from constructed diversity across SNR levels, noise backgrounds, and species co-occurrences, stabilizes the optimization trajectory towards SNR-invariant features, providing the theoretical basis for generalization to passive monitoring environments under low-resource constraints.

4 Experimental Setup

4.1 Dataset

We evaluate our approach on the BirdSet benchmark (Rauch et al., 2025b), a comprehensive collection of PAM datasets covering diverse geographic regions and acoustic environments. BirdSet provides standardized train/test splits with temporal annotations for evaluating bird species classification in real-world PAM conditions. We exclude NIPS4BPlus (NBP) as it consists solely of isolated bird segments, lacking the continuous recordings required for noise mining. Table 2 summarizes the dataset statistics for each environment.

Table 2: BirdSet evaluation dataset statistics. Train and Test columns show number of 5-second segments. #Annot indicates total annotations, J represents class imbalance, and #C denotes number of species classes per environment.

Environment	Train	Test	#Annot	J	#C
PER	16,802	15,120	14,798	0.78	132
NES	16,117	24,480	6,952	0.76	89
UHH	3,626	36,637	59,583	0.64	25
HSN	5,460	12,000	10,296	0.54	21
SSW	28,403	205,200	50,760	0.77	81
SNE	19,390	23,756	20,147	0.70	56

4.2 Evaluated Models

We evaluate SG-MSM across three categories of baselines to comprehensively benchmark its performance and validate its core components:

Standard Augmentation Baselines To evaluate the effectiveness of standard methods, we train a baseline using the identical EfficientNet-B3 architecture and focal training data as our approach, but apply the comprehensive augmentation pipeline established by the BirdSet benchmark (Rauch et al., 2025b). This standard suite includes multi-label Mixup, Time Shifting, No-Call Mixing, Gain Adjustments, and spectrogram Time/Frequency masking. To systematically assess how well these augmentations address domain shift when provided with target-site background noise, we evaluate two variants:

- **BirdSet (VOX Noise):** Utilizes the VOX auxiliary dataset, which consists of background noise from soundscape recordings without bird vocalizations, for its noise mixing augmentation. This approach is implemented by Rauch et al. (2025b) to counteract covariate shift and align the focal training data with the passive soundscape test distribution.
- **BirdSet (Domain Noise):** Utilizes the identical target-site mined background noise as SG-MSM. This allows us to directly compare our targeted synthesis against standard augmentations when both are given the same site-specific noise.

Foundation Models We compare our low-resource approach against three large-scale foundation models to contextualize how much of the domain gap can be closed purely through targeted data augmentation:

- **Perch 2.0:** The state-of-the-art supervised model (van Merriënboer et al., 2025). It utilizes an EfficientNet-B3 backbone trained on over 1.5M labeled recordings.
- **BirdMAE:** The state-of-the-art self-supervised model (Rauch et al., 2025a). It employs a ViT-Large backbone pretrained on 1.7M curated audio events and is trained with an adaptation of the BirdSet augmentation pipeline. We compare against their two reported variants: *Fine-Tuning (FT)*, where the full model is updated, and *Prototypical Probing (PP)*, which trains a lightweight prototype head on frozen features.
- **Perch 1.0:** An established benchmark for large-scale avian classification (Ghani et al., 2023). This model utilizes an EfficientNet-B1 backbone trained on 860K Xeno-Canto recordings.

Ablations To quantify the specific contributions of our framework components, we evaluate four internal baselines using the identical EfficientNet-B3 architecture:

- **w/o SG-MSM:** Trained strictly on the scarce, domain-specific focal recordings with no SG-MSM applied.

- **SG-MSM w/o SG-Mix:** Replaces the SG-Mix algorithm with random Beta$(1, 1)$ mixing applied to the mined noise, generating stochastic variants rather than targeted SNR categories while retaining the three-stage structure.
- **Stage 1 Only:** Applies only Multi-SNR Domain NoiseMix, isolating the impact of SNR categorization without overlapping or splicing.
- **Stage 1 + Stage 2:** Applies Multi-SNR Domain NoiseMix and Iterative Multi-Label Overlapping, isolating the impact of simulated temporal co-occurrence prior to spectrogram splicing.

4.3 Implementation Details

Audio Preprocessing Audio was processed at a 32 kHz sample rate. Mel spectrograms were generated using 129 mel bands, an `n_fft` of 2048, a `hop_length` of 512, $f_{\min} = 100$, and $f_{\max} = 15000$. We utilized the `noisereduce` library for stationary noise reduction, following the methodology in Sainburg et al. (2020).

Model Training All models were trained for 30 epochs using an AdamW optimizer and a cosine annealing learning rate schedule; the checkpoint with the best performance is reported. For the main SG-MSM model, we used an initial learning rate of 9×10^{-4} (decaying to 9×10^{-6}) and a batch size of 64. The w/o SG-MSM and SG-MSM w/o SG-Mix variants were trained with an initial learning rate of 1×10^{-3} (decaying to 1×10^{-5}) and a batch size of 96.

SG-MSM Hyperparameters The synthesis framework was governed by three main hyperparameters. For Stage 1, the number of SNR categories $|c|$ determines the batch expansion factor. We used $|c| = 4$ with $c \in \{S_1, S_2, S_3, S_4\}$. Since passive recordings predominantly contain low-to-medium SNR vocalizations, reduced configurations (e.g., $|c| = 3$ with $c \in \{S_2, S_3, S_4\}$) were found to provide equivalent performance, suitable for memory-constrained training. For Stage 2 (Iterative Multi-Label Overlapping), the overlap probability p_{overlap} was set to 0.9 and the maximum species $n_{\max}$ was 4. For Stage 3, the Cross-SNR Spectrogram Splicing probability p_{csc} was set to 0.9.

BirdSet Configurations For the trained BirdSet baselines, we implemented the augmentation pipeline using the exact hyperparameters defined in the official BirdSet codebase (Rauch et al., 2025b).

5 Results and Discussion

We evaluate models using three complementary metrics following the BirdSet protocol (Rauch et al., 2025b). Following Van Merriënboer et al. (2024), we consider ROC-AUC the most stable and informative metric for bioacoustics evaluation and prioritize it in our analysis. Other metrics are described in Supplementary B.

5.1 Comparison with Standard Augmentations

Table 3 details the performance of SG-MSM against the standard augmentation pipelines. SG-MSM consistently outperforms both baseline variants, achieving a mean ROC-AUC of **0.884** compared to the standard pipeline's 0.852.

Compared to the BirdSet (VOX Noise) baseline, SG-MSM achieves relative improvements of **+3.8%** in mean ROC-AUC (0.852 to 0.884), **+7.1%** in mean cmAP (0.310 to 0.332), and **+16.2%** in mean Top-1 Accuracy (0.499 to 0.580). To verify that these improvements stem from our methodology rather than simply the inclusion of target-site data, we evaluate the baseline using our identical mined background noise. When the standard pipeline utilizes site-specific noise (*Domain Noise*) rather than generic VOX noise, overall performance degrades. With the exception of the HSN domain, scores drop across all environments, reducing the mean ROC-AUC to 0.807. This indicates that merely injecting target-site noise via standard techniques like Mixup is insufficient to overcome the focal-to-passive shift. SG-MSM's ability to improve performance using this exact same noise demonstrates that the precise SNR guidance provided by SG-Mix is required to utilize target-site noise for accurate domain alignment while preserving multi-label integrity.

Table 3: Performance comparison against Standard Augmentations. All models utilize an EfficientNet-B3 backbone trained on identical focal data. Best scores are in **bold**, second-best are underlined.

Model	PER	HSN	SSW	SNE	NES	UHH	Mean
			ROC-AUC				
BirdSet (VOX Noise)	0.755	0.864	0.934	0.834	0.902	0.825	0.852
BirdSet (Domain Noise)	0.700	0.878	0.916	0.743	0.840	0.766	0.807
SG-MSM (Ours)	**0.788**	**0.934**	**0.946**	**0.891**	**0.913**	**0.830**	**0.884**
			cmAP				
BirdSet (VOX Noise)	0.221	0.431	**0.393**	0.249	0.328	**0.238**	0.310
BirdSet (Domain Noise)	0.182	0.449	0.276	0.229	0.183	0.202	0.254
SG-MSM (Ours)	**0.262**	**0.473**	0.378	**0.324**	**0.334**	0.222	**0.332**
			Top-1 Accuracy				
BirdSet (VOX Noise)	0.448	0.588	0.564	0.544	**0.493**	0.357	0.499
BirdSet (Domain Noise)	0.395	0.621	0.469	0.655	0.283	0.316	0.457
SG-MSM (Ours)	**0.584**	**0.631**	**0.625**	**0.730**	0.484	**0.423**	**0.580**

Labeled focal examples (N_f): PER (16.8K), HSN (5.5K), SSW (28.4K), SNE (19.4K), NES (16.1K), UHH (3.6K)

5.2 Comparison with Foundation Models

We present the comparative results of our framework against state-of-the-art foundation models in Table 5. Based on these results, we observe that SG-MSM achieves a mean ROC-AUC of **0.884**, proving highly competitive with the state-of-the-art supervised Perch 2.0 (0.904) and self-supervised BirdMAE (0.880) models, while surpassing Perch 1.0 (0.827) across all domains. Notably, in the HSN environment, our model achieves 0.934 ROC-AUC with only 5.5K training examples. Furthermore, on the UHH dataset—our smallest training set with only 3.6K examples—SG-MSM achieves 0.830 ROC-AUC, outperforming the massive ViT-Large based BirdMAE (0.820) despite using orders of magnitude less data and compute (Table 4).

While BirdMAE-L (FT) achieves the highest mean cmAP (0.400) and Perch 2.0 leads in mean Top-1 Accuracy (0.655) and mean ROC-AUC, SG-MSM demonstrates remarkable efficiency. By requiring only domain-specific focal recordings and unlabeled passive audio for background noise mining, SG-MSM provides a resource-efficient alternative to massive scaling.

Computational Comparison To contextualize the resource requirements of SG-MSM relative to foundation models, Table 4 compares backbone capacity and approximate computational cost. While Table 4 details encoder costs, the choice of classifier head significantly impacts total overhead. The ProtoPNet heads utilized by Perch 2.0 and BirdMAE operate on unpooled spatial feature maps, incurring a larger peak memory footprint and computational load than the standard linear head used in SG-MSM. Ultimately, this demonstrates that highly effective data augmentation can empower relatively simple, standard architectures to achieve comparable performance levels.

Table 4: Comparison of architecture and computational cost. Head parameters and FLOPs for domain-specific models (BirdMAE-L and SG-MSM) assume the PER dataset configuration ($C = 132$ classes, $J = 20$ prototypes); foundational model metrics reflect their reference configurations.

Model	**Backbone**	**Backbone Params**	**Backbone FLOPs**	**Head Type**	**Head Params**	**Head FLOPs**
Perch 2.0	EfficientNet-B3	10.7 M	1.8 G	ProtoPNet	90.9 M	2.73 G
Perch 1.0	EfficientNet-B1	6.5 M	0.7 G	Linear	14.0 M	0.03 G
BirdMAE-L	ViT-Large/16	307 M	167.6 G	ProtoPNet	2.71 M	0.34 G
SG-MSM (Ours)	EfficientNet-B3	10.7 M	1.8 G	Linear	0.20 M	0.4 MFLOPs

Table 5: Performance comparison against Foundation Models. Perch 2.0 pretrained on 1.5M recordings; Perch 1.0 on 860K recordings; BirdMAE on 1.7M events. **FT**: Fine-Tuning, **PP**: Prototypical Probing. Best scores are in **bold**, second-best are underlined.

Model	**Train**	**PER**	**HSN**	**SSW**	**SNE**	**NES**	**UHH**	**Mean**
				ROC-AUC				
Perch 2.0	1.5M	0.786	0.915	**0.973**	0.883	**0.953**	**0.912**	**0.904**
Perch 1.0	860K	0.700	0.860	0.910	0.830	0.900	0.760	0.827
BirdMAE-L (FT)	1.7M	**0.820**	0.900	0.930	0.880	0.910	0.820	0.877
BirdMAE-L (PP)	1.7M	**0.820**	0.900	0.940	0.860	0.930	0.830	0.880
SG-MSM (Ours)	[3.6K-28K]	0.788	**0.934**	0.946	**0.891**	0.913	0.830	0.884
				cmAP				
Perch 2.0	1.5M	0.232	0.533	**0.469**	**0.341**	0.403	**0.380**	0.393
Perch 1.0	860K	0.180	0.450	0.280	0.290	0.390	0.270	0.310
BirdMAE-L (FT)	1.7M	**0.350**	**0.550**	0.410	0.340	**0.410**	0.340	**0.400**
BirdMAE-L (PP)	1.7M	0.310	0.490	0.380	0.300	0.380	0.300	0.360
SG-MSM (Ours)	[3.6K-28K]	0.262	0.473	0.378	0.324	0.334	0.222	0.332
				Top-1 Accuracy				
Perch 2.0	1.5M	0.535	**0.659**	**0.789**	**0.791**	0.562	**0.595**	**0.655**
Perch 1.0	860K	0.480	0.580	0.620	0.690	**0.660**	0.570	0.600
BirdMAE-L (FT)	1.7M	**0.600**	0.640	0.700	0.610	0.520	0.420	0.582
BirdMAE-L (PP)	1.7M	0.590	0.380	0.620	0.540	0.470	0.360	0.493
SG-MSM (Ours)	[3.6K-28K]	0.584	0.631	0.625	0.730	0.484	0.423	0.580

Labeled focal examples (N_f): PER (16.8K), HSN (5.5K), SSW (28.4K), SNE (19.4K), NES (16.1K), UHH (3.6K)

5.3 Ablation Study

To show the contribution of each component within our framework, we perform an ablation study using the internal baselines previously defined in Section 4.2. We incrementally evaluate the performance of our model by isolating the individual impact of our three stages (Multi-SNR Domain NoiseMix, Iterative Multi-Label Overlapping, and Cross-SNR Spectrogram Splicing). The results are detailed in Table 6.

As shown in the table, w/o SG-MSM baseline confirms that naive domain-specific training is insufficient to overcome the focal-to-passive domain shift under low-resource constraints. This performance gap validates the field's general preference for large-scale pretraining, as evidenced by Perch 1.0 (0.827 ROC-AUC) significantly outperforming the naive baseline (Ghani et al., 2023; van Merriënboer et al., 2025). However, applying SG-MSM to the *same* low-resource data and simple architecture as the ablation lifts mean ROC-AUC from 0.751 to 0.884, representing a **+17.7%** relative gain. This transformation is consistent across metrics, with

relative gains of **+68.5%** in cmAP and **+69.1%** in Top-1 Accuracy. These results suggest that SG-MSM effectively bridges the domain gap by transforming weak, domain-specific baselines into robust classifiers, offering a highly effective strategy for data-constrained deployments. The consistent improvement over the ablation baseline confirms SG-MSM as the primary driver of performance.

SG-MSM provides consistent improvements over SG-MSM w/o SG-Mix across all mean metrics: ROC-AUC (**+3.0%**), cmAP (**+7.4%**), and Top-1 Accuracy (**+29.2%**) relative gains. Even with the same three-stage framework structure, SNR-guided mixing significantly outperforms random Beta$(1, 1)$ mixing, validating that precise SNR control is important for domain adaptation. The per-domain improvements are substantial: HSN shows the largest ROC-AUC improvement (**+8.7%** relative gain), indicating SG-Mix is especially beneficial for smaller, challenging datasets. This ablation confirms that the SG-Mix algorithm itself is a core contribution of the multi-stage framework.

To isolate the contribution of each component, we evaluate the pipeline incrementally. Applying only Multi-SNR Domain NoiseMix (Stage 1) improves the baseline mean ROC-AUC from 0.751 to 0.772. The addition of Iterative Multi-Label Overlapping (Stage 2) increases performance to 0.859, highlighting the role of simulated species co-occurrence. Finally, integrating Cross-SNR Spectrogram Splicing (Stage 3) yields the full framework's 0.884 ROC-AUC. These results confirm that while multi-label overlapping provides the largest absolute gain, all three stages collectively improve performance across all evaluated metrics.

Table 6: Ablation Study isolating the contribution of SG-MSM components. Best scores are in **bold**, second-best are underlined.

Ablation	PER	HSN	SSW	SNE	NES	UHH	Mean
			ROC-AUC				
w/o SG-MSM	0.673	0.732	0.800	0.708	0.839	0.756	0.751
SG-MSM w/o SG-Mix	0.785	0.859	0.943	0.849	0.904	0.810	0.858
Stage 1 Only	0.643	0.824	0.820	0.781	0.830	0.734	0.772
Stage 1 + Stage 2	0.770	0.892	0.942	0.857	0.887	0.808	0.859
SG-MSM (Full)	**0.788**	**0.934**	**0.946**	**0.891**	**0.913**	**0.830**	**0.884**
			cmAP				
w/o SG-MSM	0.124	0.203	0.227	0.203	0.265	0.158	0.197
SG-MSM w/o SG-Mix	**0.265**	0.386	**0.379**	0.301	0.290	**0.230**	0.309
Stage 1 Only	0.153	0.411	0.334	0.240	0.289	0.149	0.263
Stage 1 + Stage 2	0.243	**0.484**	0.366	0.258	0.309	0.200	0.310
SG-MSM (Full)	0.262	0.473	0.378	**0.324**	**0.334**	0.222	**0.332**
			Top-1 Accuracy				
w/o SG-MSM	0.340	0.256	0.440	0.345	0.378	0.298	0.343
SG-MSM w/o SG-Mix	0.566	0.461	0.414	0.585	0.364	0.304	0.449
Stage 1 Only	0.442	0.589	0.588	0.563	0.409	0.356	0.491
Stage 1 + Stage 2	0.501	0.619	0.545	0.575	0.446	0.403	0.515
SG-MSM (Full)	**0.584**	**0.631**	**0.625**	**0.730**	**0.484**	**0.423**	**0.580**

Labeled focal examples (N_f): PER (16.8K), HSN (5.5K), SSW (28.4K), SNE (19.4K), NES (16.1K), UHH (3.6K)

6 Conclusion

We presented SG-MSM, a framework demonstrating that targeted data augmentation can rival the performance of large-scale pretraining for domain-shifted bioacoustic tasks. Our findings show that *precise distribution alignment* offers a highly efficient alternative to massive pretraining, proving that robustness does not strictly require massive data scaling. By effectively bridging the domain gap with orders of magnitude less data, SG-MSM offers a democratized path for deploying high-performance PAM models in resource-constrained environments where large-scale pretraining is inaccessible. Furthermore, the model-agnostic nature of our framework suggests that SG-MSM could serve as a complementary augmentation strategy for adapting future bioacoustic foundation models to passive environments, ensuring that high-performance classification remains adaptable to the complex, diverse soundscapes of the real world.

Finally, uniting the evident power of scaling (Kaplan et al., 2020) with SG-MSM's adaptation capabilities, we propose a high-resource roadmap: applying SG-MSM to global bioacoustic archives with diverse soundscapes. This envisions a **Global PAM Model** capable of high-performance, robust global biodiversity monitoring.

7 Limitations and Generalizability

Limitations Unlike foundation models (e.g., Perch 2.0) that offer *out-of-the-box inference* capabilities for thousands of species and robust few-shot transfer to novel classes, SG-MSM—in its current low-resource implementation—acts strictly as a *domain-specific data augmentation framework*.

Generalizability While SG-MSM was evaluated on avian bioacoustics, its core mechanism addresses the universal challenge of SNR-driven distribution shifts in multilabel PAM. We posit that the framework's principles are taxon-agnostic, suggesting potential for broader application in domains like marine bioacoustics (e.g., cetacean monitoring) or industrial audio analysis, where similar focal-to-passive discrepancies exist.

References

Hicham Bellafkir, Markus Vogelbacher, Daniel Schneider, Markus Mühling, Nikolaus Korfhage, and Bernd Freisleben. Edge-based bird species recognition via active learning. In *International Conference on Networked Systems*, pp. 17–34. Springer, 2023.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.

Malik Boudiaf, Tom Denton, Bart Van Merriënboer, Vincent Dumoulin, and Eleni Triantafillou. In search for a generalizable method for source free domain adaptation. In *International Conference on Machine Learning*, pp. 2914–2931. PMLR, 2023.

Juan Sebastián Cañas, Stefan Kahl, Tom Denton, María Paula Toro-Gómez, Susana Rodriguez-Buritica, Jose Luis Benavides-Lopez, Juan Sebastián Ulloa, Paula Caycedo-Rosales, Holger Klinck, Hervé Goëau, et al. Overview of birdclef+ 2025: Multi-taxonomic sound identification in the middle magdalena, colombia. In *CLEF 2025-Working Notes of the Conference and Labs of the Evaluation Forum*, volume 4038, pp. 2909–2919. CEUR, 2025.

SM Chowfin and AJ Leslie. Using birds as biodindicators of forest restoration progress: A preliminary study. *Trees, Forests and People*, 3:100048, 2021.

Tom Denton, Scott Wisdom, and John R Hershey. Improving bird classification with unsupervised sound separation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 636–640. IEEE, 2022.

Burooj Ghani, Tom Denton, Stefan Kahl, and Holger Klinck. Global birdsong embeddings enable superior transfer learning for bioacoustic classification. *Scientific Reports*, 13(1):22876, 2023.

Jenny Hamer, Eleni Triantafillou, Bart Van Merriënboer, Stefan Kahl, Holger Klinck, Tom Denton, and Vincent Dumoulin. Birb: A generalization benchmark for information retrieval in bioacoustics. *arXiv preprint arXiv:2312.07439*, 2023.

René Heinrich, Lukas Rauch, Bernhard Sick, and Christoph Scholz. Audioprotopnet: An interpretable deep learning model for bird sound classification. *Ecological Informatics*, 87:103081, 2025.

Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, et al. Cnn architectures for large-scale audio classification. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 131–135, 2017.

Elad Hoffer, Tal Ben-Nun, Itay Hubara, Niv Giladi, Torsten Hoefler, and Daniel Soudry. Augment your batch: Improving generalization through instance repetition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8129–8138, 2020.

Stefan Kahl, Connor M Wood, Maximilian Eibl, and Holger Klinck. Birdnet: A deep learning solution for avian diversity monitoring. *Ecological Informatics*, 61:101236, 2021.

Stefan Kahl, Tom Denton, Holger Klinck, Vijay Ramesh, Viral Joshi, Meghana Srivathsa, Akshay Anand, Chiti Arvind, Harikrishnan Cp, Suyash Sawant, et al. Overview of birdclef 2024: Acoustic identification of under-studied bird species in the western ghats. In *CLEF 2024-25th Conference and Labs of the Evaluation Forum*, pp. 1948–1957, 2024.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

V. Nolan, C. Scott, J. M. Yeiser, N. Wilhite, P. E. Howell, D. Ingram, and J. A. Martin. The development of a convolutional neural network for the automatic detection of northern bobwhite (*colinus virginianus*) covey calls. *Remote Sensing in Ecology and Conservation*, 9(1):46–61, 2023.

Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*, 2019.

Vijay Ramesh, Priyanka Hariharan, VA Akshay, Pooja Choksi, Sarika Khanwilkar, Ruth DeFries, and VV Robin. Using passive acoustic monitoring to examine the impacts of ecological restoration on faunal biodiversity in the western ghats. *Biological Conservation*, 282:110071, 2023.

Ezhini Rasendiran R and Chandresh Kumar Maurya. Improving Bird Classification with Primary Color Additives. In *Interspeech 2025*, pp. 1703–1707, 2025. doi: 10.21437/Interspeech.2025-2516.

Lukas Rauch, Denis Huseljic, Moritz Wirth, Jens Decke, Bernhard Sick, and Christoph Scholz. Towards deep active learning in avian bioacoustics, 2024.

Lukas Rauch, René Heinrich, Ilyass Moummad, Alexis Joly, Bernhard Sick, and Christoph Scholz. Can masked autoencoders also listen to birds? *arXiv preprint arXiv:2504.12880*, 2025a.

Lukas Rauch, Raphael Schwinger, Moritz Wirth, René Heinrich, Denis Huseljic, Marek Herde, Jonas Lange, Stefan Kahl, Bernhard Sick, Sven Tomforde, and Christoph Scholz. Birdset: A large-scale dataset for audio classification in avian bioacoustics. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=dRXxFEY8ZE.

Tim Sainburg and Asaf Zorea. Domain general noise reduction for time series signals with noisereduce. *Scientific Reports*, 15(1):30905, 2025.

Tim Sainburg, Marvin Thielk, and Timothy Q Gentner. Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires. *PLoS computational biology*, 16(10):e1008228, 2020.

Xiaona Sun, Zhenyu Wu, Zhiqiang Zhan, and Yang Ji. Contrastive conditional alignment based on label shift calibration for imbalanced domain adaptation. In *International Conference on Pattern Recognition*, pp. 13–28. Springer, 2025.

Richard Sutton. The bitter lesson. *Incomplete Ideas (blog)*, 13(1):38, 2019.

Bhuvaneswari Swaminathan, M. Jagadeesh, and Subramaniyaswamy Vairavasundaram. Multi-label classification for acoustic bird species detection using transfer learning approach. *Ecological Informatics*, 80: 102471, 2024. ISSN 1574-9541.

Mingxing Tan and Quoc V Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 6105–6114. PMLR, 2019.

Bart Van Merriënboer, Jenny Hamer, Vincent Dumoulin, Eleni Triantafillou, and Tom Denton. Birds, bats and beyond: Evaluating generalization in bioacoustics models. *Frontiers in Bird Science*, 3:1369756, 2024.

Bart van Merriënboer, Vincent Dumoulin, Jenny Hamer, Lauren Harrell, Andrea Burns, and Tom Denton. Perch 2.0: The bittern lesson for bioacoustics. *arXiv preprint arXiv:2508.04665*, 2025.

Scott Wisdom, Efthymios Tzinis, Hakan Erdogan, Ron Weiss, Kevin Wilson, and John Hershey. Unsupervised sound separation using mixture invariant training. *Advances in neural information processing systems*, 33:3846–3857, 2020.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

A Detailed Algorithms

This appendix provides the full pseudo-code for the proposed SG-MSM framework (Algorithms 3, 4, and 5).

Algorithm 3 employs iterative noisereduce-based SNR calculation with binary search optimization and boundary SNR caching. The algorithm internally verifies a 10 dB SNR gap between focal recording and background noise to prevent signal masking; if needed, the focal audio undergoes self-noise-reduction enhancement ($x_{\text{signal}} = \texttt{noisereduce}(x_f)$, $x_{\text{noise}} = x_f - x_{\text{signal}}$, then $x_f \leftarrow \text{normalize}(\alpha \cdot x_{\text{signal}} + (1-\alpha) \cdot x_{\text{noise}})$ with $\alpha \in [0.75, 0.95]$ until the gap is achieved).

Algorithm 3 provides accurate SNR control through iterative noisereduce-based calculation with binary search optimization. To address the computational expense of iterative noisereduce calls, Algorithm 5 introduces (1) signal/noise component caching and (2) linear interpolation correction. The accelerated approach pre-computes signal and noise components once, then uses cached linear combinations for rapid SNR estimation, followed by single-point accuracy correction through linear interpolation to precisely achieve target SNR levels. This hybrid strategy achieves approximately 3× speedup with 0.3dB precision trade-off compared to Algorithm 3, making it suitable for large-scale training scenarios.

Algorithm 3 SG-Mix

Input: Focal recording x_f, background noise b, target SNR μ, tolerance ϵ
Output: Mixed audio x_{mix} with target SNR

Function CalculateSNR(x_{mix}):
  $x_{denoised} \leftarrow$ noisereduce(x_{mix}, stationary=True)
  $x_{noise} \leftarrow x_{mix} - x_{denoised}$
  $P_{signal} \leftarrow \frac{1}{N}\sum_{i=1}^{N} x_{denoised}[i]^2$
  $P_{noise} \leftarrow \frac{1}{N}\sum_{i=1}^{N} x_{noise}[i]^2$
  return $10\log_{10}(P_{signal}/P_{noise})$

Function MixAndCalculateSNR(α):
  $x_{mix} \leftarrow \alpha \cdot x_f + (1-\alpha)\cdot b$
  $x_{mix} \leftarrow x_{mix}/\max(|x_{mix}|)$ {Normalize}
  return CalculateSNR(x_{mix})

Binary Search for Optimal Mixing Ratio:
$\alpha_{left} \leftarrow 0.0$, $\alpha_{right} \leftarrow 1.0$
$\alpha_{best} \leftarrow 0.0$, $diff_{min} \leftarrow \infty$
$SNR_{left} \leftarrow$ None, $SNR_{right} \leftarrow$ None {Caching optimization}
for $iter = 1$ to max_iter {set to 30 in our case} **do**
  $\alpha_{mid} \leftarrow (\alpha_{left} + \alpha_{right})/2$
  if SNR_{left} is None **then**
    $SNR_{left} \leftarrow$ MixAndCalculateSNR(α_{left})
  end if
  if SNR_{right} is None **then**
    $SNR_{right} \leftarrow$ MixAndCalculateSNR(α_{right})
  end if
  $SNR_{mid} \leftarrow$ MixAndCalculateSNR(α_{mid}) {Always new}
  Update α_{best} with α giving minimum $|SNR - \mu|$
  if $|SNR_{best} - \mu| \leq \epsilon$ **then**
    break {Acceptable tolerance reached}
  end if
  if $SNR_{mid} > \mu$ **then**
    $\alpha_{right} \leftarrow \alpha_{mid}$, $SNR_{right} \leftarrow SNR_{mid}$ {Reuse}
  else
    $\alpha_{left} \leftarrow \alpha_{mid}$, $SNR_{left} \leftarrow SNR_{mid}$ {Reuse}
  end if
end for
$x_{mix} \leftarrow \alpha_{best} \cdot x_f + (1-\alpha_{best})\cdot b$
return $x_{mix}/\max(|x_{mix}|)$

Algorithm 4 Iterative Multi-Label Overlapping

Input: SNR variant $x_i^{(c)}$ from Stage 1, overlap probability p_{overlap}, maximum species $n_{\max}$, margin $\delta = 0.2$ dB

Output: Multi-species overlapped audio $x_i^{(c)}{}_{\text{final}}$ with unified labels y_{final}

if random() $> p_{\text{overlap}}$ **then**
  return $x_i^{(c)}$, y_i {Skip overlapping}
end if

$n_{\text{species}} \leftarrow \text{random}\{2, 3, \ldots, n_{\max}\}$ {Total species desired}
$x_{\text{current}} \leftarrow x_i^{(c)}$, $y_{\text{current}} \leftarrow y_i$
$\text{SNR}_{\text{current}} \leftarrow \text{CalculateSNR}(x_{\text{current}})$

For first category (S1): {Generate and cache secondary examples}
SecondaryCache $\leftarrow$ [ ] {Empty list for caching}
for $r = 1$ to $n_{\text{species}} - 1$ **do**
  Select secondary example x_j meeting constraints:
    (1) $\text{id_snr}_j > \text{SNR}_{\text{current}}$ {Higher quality}
    (2) NOT ($\text{id_snr}_j \leq -15$ AND $\text{id_entropy}_j \leq 0.35$) {Exclude low quality}
    (3) Random peak-centered 5-second segment from x_j

  SecondaryCache.append(x_j, y_j) {Cache for other categories}

  $\text{SNR}_{\text{target}} \leftarrow \text{SNR}_{\text{current}} + \delta$ {Small margin increase}
  $x_{\text{current}} \leftarrow \text{SG-Mix}(x_j, x_{\text{current}}, \text{SNR}_{\text{target}})$
  $y_{\text{current}} \leftarrow y_{\text{current}} \cup y_j$ {Union labels}
  $\text{SNR}_{\text{current}} \leftarrow \text{CalculateSNR}(x_{\text{current}})$ {Update for next constraint}
end for

For other categories (S2, S3, S4): {Reuse cached secondary examples}
for each (x_j, y_j) in SecondaryCache **do**
  $\text{SNR}_{\text{target}} \leftarrow \text{SNR}_{\text{current}} + \delta$
  $x_{\text{current}} \leftarrow \text{SG-Mix}(x_j, x_{\text{current}}, \text{SNR}_{\text{target}})$
  $y_{\text{current}} \leftarrow y_{\text{current}} \cup y_j$
  $\text{SNR}_{\text{current}} \leftarrow \text{CalculateSNR}(x_{\text{current}})$
end for

return x_{current}, y_{current}

Algorithm 5 Accelerated SG-Mix

Input: Focal recording x_f, background noise b, target SNR μ, tolerance ϵ
Output: Mixed audio x_{mix} with target SNR

Function CalculateSNRCached(x_{signal}, x_{noise}):
  $P_{signal} \leftarrow \frac{1}{N}\sum_{i=1}^{N} x_{signal}[i]^2$
  $P_{noise} \leftarrow \frac{1}{N}\sum_{i=1}^{N} x_{noise}[i]^2$
  return $10\log_{10}(P_{signal}/P_{noise})$

Signal/noise component caching:
$x_{f_signal} \leftarrow$ noisereduce(x_f, stationary=True)
$x_{f_noise} \leftarrow x_f - x_{f_signal}$
$x_{b_signal} \leftarrow$ noisereduce(b, stationary=True)
$x_{b_noise} \leftarrow b - x_{b_signal}$

Function MixAndCalculateSNRCached(α):
  $x_{mixed_signal} \leftarrow \alpha \cdot x_{f_signal} + (1-\alpha) \cdot x_{b_signal}$
  $x_{mixed_noise} \leftarrow \alpha \cdot x_{f_noise} + (1-\alpha) \cdot x_{b_noise}$
  Normalize by total mixed audio magnitude
  return CalculateSNRCached(x_{mixed_signal}, x_{mixed_noise})

Binary Search for Optimal Mixing Ratio:
$\alpha_{cached} \leftarrow$ Binary search procedure with MixAndCalculateSNRCached(α)
$SNR_{cached} \leftarrow$ MixAndCalculateSNRCached(α_{cached})

if $|SNR_{cached} - \mu| \leq 0.05$ **then**
  return α_{cached} {Early exit if sufficient accuracy}
end if

Linear interpolation correction:
$\alpha_{sample} \leftarrow \alpha_{cached} \pm 0.025$ {Sign matches SNR error sign}
$SNR_{sample} \leftarrow$ AccurateSNR(α_{sample}) {One noisereduce call}
$\alpha_{final} \leftarrow$ LinearInterpolate(α_{cached}, α_{sample}, SNR_{cached}, SNR_{sample}, μ)
$x_{mix} \leftarrow \alpha_{final} \cdot x_f + (1-\alpha_{final}) \cdot b$
return $x_{mix}/\max(|x_{mix}|)$

B Evaluation Metrics

We evaluate models using three complementary metrics following the BirdSet protocol (Rauch et al., 2025b). **ROC-AUC** (Area Under the Receiver Operating Characteristic Curve) measures the probability that a randomly chosen positive instance is ranked higher than a randomly chosen negative instance, providing threshold-independent assessment balanced across class distributions. **cmAP** (class-mean Average Precision) computes average precision per class and averages across all classes, ensuring equal weighting regardless of class imbalance. **Top-1 Accuracy** evaluates whether the highest-confidence prediction matches any true label; although less meaningful for multilabel instances, it provides a quick assessment of the model's ability to correctly identify at least one relevant class. Following Van Merriënboer et al. (2024), we consider ROC-AUC the most stable and informative metric for bioacoustics evaluation and prioritize it in our analysis.