Comparative Analysis of Visual and Acoustic Deep Learning Models for Bird Classification

Anne Carvalho [*1] **Giovanni Martinelli** [*1] **Lucas Andrade** [*1] **Victor Augusto L. Cruz** [*1]

Abstract

The accurate identification of birds and other morphologically similar animals is essential in biodiversity monitoring and conservation contexts, as it underpins processes such as population estimation, geographic distribution mapping, and the detection of population declines. Misclassification can compromise ecological assessments and negatively influence management and conservation policies. Moreover, properly distinguishing between species with similar appearances is crucial for mitigating ecological impacts, since visually nearly identical species may play very different roles in the ecosystem, acting, for example, as disease vectors, invasive species, or key native species. Therefore, in this work, different methodologies for bird species identification are explored, aiming to uncover the best attributes for such tasks. The findings presented here demonstrate the superiority of Transformer-based architectures in both visual and acoustic domains. Specifically, Vision Transformers (ViT) achieved 92.33% accuracy on the audiovisual intersection subset, while Audio Spectrogram Transformers (AST) using Log-Mel Spectrograms reached 57.28%, significantly outperforming convolutional baselines and demonstrating the efficacy of attention mechanisms for fine-grained categorization.

1. Introduction

Image classification is a central task in the field of computer vision, being defined as the process of assigning a categorical label to an input image based on its visual properties and the spatial patterns it contains. In the early stages of the field, this process relied mainly on handcrafted feature extraction methods, such as Histogram of Oriented Gradients (HOG) and Scale-Invariant Feature Transform (SIFT). Although effective in a variety of scenarios, such methods exhibited limitations stemming from their rigid and heuristic-driven nature and their limited ability to generalize to complex variations in images.

The advent of Convolutional Neural Networks (CNNs) introduced a significant paradigm shift by enabling discriminative representations to be learned directly from data. This approach allowed models to learn progressively more abstract and hierarchically structured features, resulting in substantial performance gains. Architectures such as ResNet (He et al., 2016), DenseNet, and EfficientNet have become established state-of-the-art references for classification tasks in recent years. More recently, however, Vision Transformers (ViT) have gained prominence by adapting self-attention mechanisms—originally developed for Natural Language Processing—to the vision domain. These architectures have demonstrated performance competitive with traditional CNNs, especially in large-scale settings.

The continuous advancement of these models has enabled the application of deep learning techniques in specialized domains, such as bird species classification, the central theme of this work. Automatic bird identification plays a key role in ecological and biological studies, contributing to several tasks, such as ecosystem monitoring, biodiversity conservation, and the mitigation of illegal hunting practices. However, bird classification presents challenges that differ substantially from those encountered in conventional image classification tasks. In particular, biodiversity datasets often exhibit severe class imbalance due to the greater abundance of common species compared to the limited availability of images for rare species. Furthermore, the task is characterized as a Fine-Grained Image Classification problem, marked by high intra-class variability—stemming from variations in pose, illumination, plumage, and occlusions—and low inter-class variability, as morphological differences between distinct species may be visually subtle. These factors require models capable of capturing extremely fine visual nuances while remaining robust to significant internal variations.

[*]Equal contribution [1]Department of Computer Science, Universidade Federal de Minas Gerais, 31270-901, Belo Horizonte, MG, Brazil.. Correspondence to: Giovanni Martinelli <giovanni@ufmg.br>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

Furthermore, biological classification offers unique multimodal opportunities; specifically, vocalizations provide a distinct signal orthogonal to visual morphology. So, in this research we want to try and use both the visual aspect and the auditory aspect on the Fine-Grained Image Classification problem to try and identify what are the more defining features on birds on the computational point of view.

2. Related Work

Several studies have been proposed to address the challenges of bird species classification. Islam et al. (2019), in Bird Species Classification from an Image Using VGG-16 Network, employ transfer learning using the VGG-16 architecture as a feature extractor. From the resulting 4096-dimensional feature vectors, different supervised classifiers such as Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Random Forest (RF) were evaluated. The linear-kernel SVM achieved the best performance, reaching an accuracy of 89%.

Anusha & ManiSai (2022), in Bird Species Classification Using Deep Learning, adopt an approach based on Deep Convolutional Neural Networks, aiming to capture distinctive visual features such as beak shape, plumage patterns, and head region. Using the CUB-200-2011 dataset, their method achieved 89% accuracy, surpassing earlier benchmark models such as Pose Normalization (82%) and Multiple-Granularity CNN (83%).

In another direction, Marini et al. (2013), in Bird Species Classification Based on Color Features, investigate an approach based exclusively on color information, applying chromatic segmentation to isolate the background and constructing histograms in RGB and HSV color spaces. A multiclass SVM was used for the classification step, resulting in a maximum accuracy of 8.60%, showing that color features alone are insufficient to distinguish a large number of species.

Finally, Kumar et al. (Das & Kumar, 2018), in Bird Species Classification Using Transfer Learning with Multistage Training, propose a multi-stage strategy combining detection and classification. First, a Mask R-CNN pre-trained on the COCO dataset is used to localize regions of interest containing the birds. Next, an ensemble composed of Inception ResNet V2 and Inception V3 is trained in two phases: (i) using the original images and (ii) using cropped images, with the goal of capturing both macro- and micro-structural features. On the Indian bird dataset, the method achieved an F1-score of 55.67%.

In the domain of audio classification, BirdNET (Kahl et al., 2021) stands as the industry standard for avian diversity monitoring, utilizing a custom ResNet architecture to identify over 3,000 species from their vocalizations. More recently, Google's Perch (Ghani et al., 2023) has pushed the state-of-the-art by employing efficient Transformer-based models for global-scale bird species recognition in real-time.

Regarding Fine-Grained Visual Classification (FGVC) specifically for the CUB-200 dataset, specialized architectures have been proposed to handle high intra-class variance. NTS-Net (Yang et al., 2018) introduces a Navigator-Teacher-Scrutinizer network to automatically locate informative regions without bounding box annotations. Similarly, TransFG (He et al., 2022) adapts the Vision Transformer architecture to capture subtle discriminative features by selecting top-k attention patches, demonstrating superior performance on fine-grained benchmarks.

3. Methodology

3.1. Datasets Characterization

This work explores bird classification using the popular CUB 200 Dataset (Wah et al., 2011), which covers 200 bird species in a total of 11,788 images. The dataset is notable for its richness in annotations, including 312 visual attributes per image, and for the high variability of the captures (different angles and scenarios). Despite its high quality and completeness, the limited number of images per class (around 60) represents a constraint for classification architectures that require large volumes of data. For processing, all images, originally of varying resolutions, were resized to 224x224 pixels.

For audio classification, we constructed a dataset from Xeno-Canto (Xeno-canto Foundation, 2025) containing recordings for the species present in CUB-200. To enable multimodal analysis, we specifically focused on the intersection of these two datasets, resulting in a subset of 90 bird species common to both CUB-200 and Xeno-Canto. This intersection dataset allows for direct comparison and future fusion of visual and acoustic modalities.

Figure 1 illustrates the distribution of audio samples across the intersection dataset. The disparity is driven by the nature of citizen science data collection: synanthropic species (those living near humans) like the House Sparrow are overrepresented, while elusive or migratory species like the Hooded Merganser have scarce recordings.

3.2. Image Classification

For the task of image classification in the context of the CUB-200-2011 dataset, the objective is to analyze the difficulties inherent to Fine-Grained Visual Classification (FGVC) and to understand the limitations and effectiveness of different architectures for bird-species identification. The methodology was divided into three main approaches.

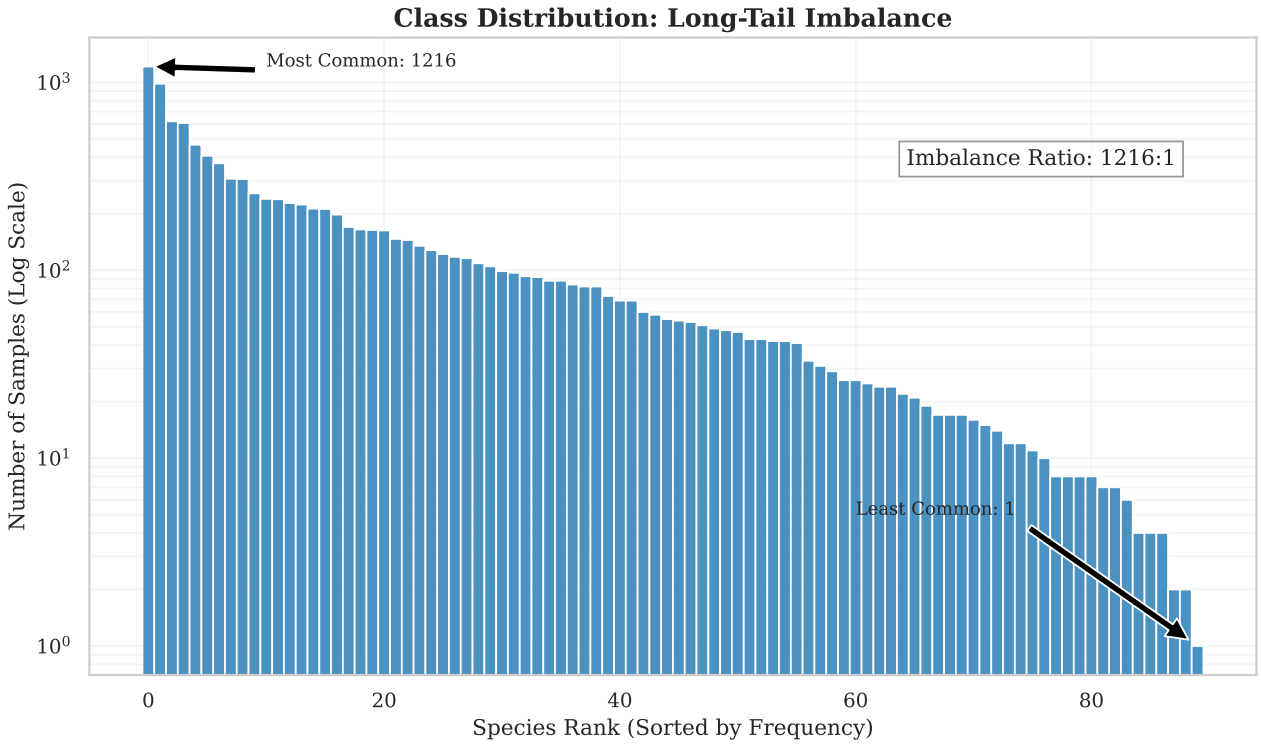

Figure 1. Log-scale distribution of sample counts per species in the raw intersection dataset. The "long-tail" imbalance (1216:1 ratio) is clearly visible. For training, species with fewer than 2 samples were filtered, reducing the effective imbalance to 608:1.

Attribute-Based Visual Classification (Traditional Approach) The first stage explored classification using exclusively the 312 visual attributes provided by the dataset. These attributes explicitly encode morphological and phenotypic properties such as beak morphology, feather coloration and patterns, tail shape, and eye color.

Gaussian Naive Bayes was first employed as a probabilistic baseline (Hand & Yu, 2001). This model assumes conditional independence among features and models each attribute using a class-specific Gaussian distribution. Although this assumption is rarely true in fine-grained visual tasks—where attributes exhibit strong correlation—it provides a lightweight generative baseline and serves as a lower bound for performance.

K-Nearest Neighbors (KNN) was evaluated to assess instance-based classification in the attribute space (Cover & Hart, 1967). Using k=5, the model assigns each sample the label shared by the majority of its nearest neighbors under Euclidean distance. KNN is sensitive to local geometry and therefore helpful for examining whether species clusters are separable based purely on attribute vectors.

Decision Trees and Random Forests were included to investigate non-linear decision boundaries. A Decision Tree with maximum depth 10 restricts memorization and promotes generalizable rules based on attribute thresholds (Breiman et al., 1984). Random Forests extend this idea by aggregating 300 randomized trees, reducing variance and improving robustness to noisy or redundant attributes, which are common in manually annotated datasets (Breiman, 2001).

Linear Discriminant Analysis (LDA) was used to measure how well linear projections separate bird species given their semantic attributes (Fisher, 1936). By maximizing between-class variance while minimizing within-class variance, LDA tests whether attribute combinations form linearly discriminative subspaces—a property relevant for FGVC, where subtle attribute differences often encode species-specific traits.

Support Vector Machines (SVMs) were then applied as margin-based classifiers, using class-weight balancing to mitigate mild imbalance in attribute distributions (Cortes & Vapnik, 1995). The SVM seeks the hyperplane that maximizes the decision margin, enabling it to capture fine-grained distinctions between species encoded in the attributes. Probability calibration was enabled to support later comparative evaluation.

Finally, Logistic Regression with the lbfgs solver and 1000 iterations provided a linear probabilistic baseline (Cox, 1958). By modeling class likelihoods through a softmax function over weighted attribute combinations, it evaluates whether the semantic attributes contain linearly separable information across all 200 species.

Each model was evaluated using 5-fold cross-validation, and results were compared in terms of accuracy and F1-score. Among these, SVM achieved the best performance (47% accuracy), confirming its ability to capture complex decision boundaries, though still limited compared to deep learning approaches.

Classification with Convolutional Neural Networks (CNNs) The second stage explored deep vision architectures, specifically EfficientNetB0 (Tan & Le, 2019), initialized with ImageNet-pretrained weights.

Because of limitations in processing time and computational resources, we adopted a fine-tuning strategy that freezes the initial convolutional layers and trains only the upper layers and classification head. This allowed us to evaluate how effectively ImageNet knowledge transfers to the fine-grained bird-classification domain.

For EfficientNetB0, the base network was loaded without the fully connected head and using global average pooling. Fine-tuning was conducted by unfreezing only the top 25 layers, while the earlier convolutional layers remained frozen to preserve low-level pretrained representations. The classification head consisted of a Batch Normalization layer, followed by Dropout (0.25), a dense projection of 512 units with ReLU activation, an additional Dropout layer (0.25), and a final softmax output.

The model was trained using the Adam optimizer, early stopping based on validation accuracy with a patience of five epochs, and a maximum of 20 training epochs. The data were split into 80% training, 10% validation, and 10% testing.

Classification with Vision Transformers (ViT) Finally, models from the Vision Transformer (ViT) family (Dosovitskiy et al., 2020) were employed, representing state-of-the-art self-attention–based architectures in computer vision. We specifically utilized the `vit-base-patch16-224` architecture. This model divides the 224x224 input image into a sequence of 16x16 patches, which are linearly projected and processed by a stack of Transformer encoders. Unlike CNNs, ViT relies on self-attention mechanisms to capture global dependencies across the entire image.

Two variants were evaluated: one pretrained on ImageNet-1k and another on ImageNet-21k, the latter providing substantially broader semantic coverage. Unlike CNNs, Vision Transformers rely less on built-in inductive biases—such as locality and translation equivariance—and therefore depend more heavily on either large-scale training datasets or strong data-augmentation schemes to achieve high generalization. Empirically, preliminary experiments indicated that ViT performance improved noticeably when exposed to more diverse training data, motivating the evaluation of multiple augmentation regimes.

The first regime used no data augmentation, applying only resizing and normalization, serving as a minimal baseline. The second regime incorporated MixUp and CutMix (Yun et al., 2019), two data-mixing techniques that blend images and labels either linearly or through spatial masking. These methods act as strong regularizers by preventing the model from overfitting to spurious fine-grained details and encouraging learning of more global, robust features. The third regime combined CutMix, MixUp, and RandAugment (Cubuk et al., 2020), adding stochastic geometric and color transformations—such as rotations, elastic distortions, and brightness or contrast perturbations—to further increase synthetic diversity in the training set.

These configurations enabled the investigation of modern regularization strategies and the benefits of large-scale pretraining for improving fine-grained species-classification accuracy.

3.3. Audio Classification

To complement the visual analysis, we investigate audio-based species identification using bird vocalizations. This modality presents unique challenges, including variable signal duration, background noise, and the need for effective time-frequency representations.

Audio Feature Engineering Initial experiments employed Mel-Frequency Cepstral Coefficients (MFCCs) (Davis & Mermelstein, 1980), a standard feature in speech recognition. We constructed a 3-channel input tensor consisting of static MFCCs stacked with their first and second derivatives (Delta and Delta-Delta) to capture temporal dynamics (Abdel-Hamid et al., 2014). However, this representation yielded suboptimal performance (approximately 38% accuracy). We observed severe training instability and convergence issues, partly attributed to the difficulty in effectively normalizing these high-variance coefficients across a diverse dataset. Furthermore, MFCCs decorrelate the signal, removing spectro-temporal structures that Convolutional Neural Networks (CNNs) and Transformers exploit effectively.

Consequently, we adopted Log-Mel Spectrograms (LMS) as the primary input representation for our best performing model (AST). LMS preserves the time-frequency locality of the signal, treating audio as an image-like tensor suit-

able for 2D architectures. We processed audio clips with a sampling rate (F_s) of 22.05 kHz. Spectrograms were generated using a Fast Fourier Transform (FFT) window size (n_fft) of 2048, a hop length of 512, and 128 Mel filter banks ($n\_mels = 128$). The resulting spectrograms were padded or truncated to a fixed length of 1024 frames, resulting in a 128x1024 input tensor. This configuration captures sufficient frequency resolution for bird calls while maintaining temporal precision.

Network Architectures We evaluated three distinct architectures adapted for audio classification, corresponding to the phases of our experimentation:

Phase 0: AudioViT (Baseline): To establish a baseline, we applied a standard Vision Transformer (ViT-Base) directly to MFCC features. The MFCCs were resized to 224x224 to match the expected input of the pretrained ViT. This naive approach served to test the transferability of visual models to cepstral audio representations.

Phase 2: AudioCNNv2: We utilized a 14-layer Convolutional Neural Network inspired by the PANNs architecture (Kong et al., 2020). This model processes MFCCs using a hierarchy of convolutional blocks with increasing channel width (up to 512), designed to capture local spectro-temporal patterns. It was trained using Focal Loss to handle class imbalance.

Phase 3: Audio Spectrogram Transformer (AST): Our final and most advanced model, AST (Gong et al., 2021), is a specialized Transformer for audio. It takes Log-Mel Spectrograms (128x1024) as input, splits them into 16x16 patches, and applies self-attention. AST is initialized with ImageNet weights and further pretrained on AudioSet, allowing it to effectively capture long-range dependencies in bird calls.

Class Imbalance Mitigation The raw dataset exhibits severe class imbalance, with a ratio of 1216:1 between the most and least common classes. However, for our experiments, we filtered out species with fewer than 2 samples, resulting in an effective imbalance of 608:1. To address this, we implemented two key strategies:

Focal Loss: We replaced the standard Cross-Entropy loss with Focal Loss (Lin et al., 2017), defined as $FL(p_t) = -\alpha_t(1-p_t)^\gamma \log(p_t)$. By setting the focusing parameter $\gamma = 2.0$, the loss function down-weights easy examples and focuses training on hard, misclassified examples, preventing the vast number of easy negatives from overwhelming the gradient.

Advanced Augmentation: We applied SpecAugment (Park et al., 2019), which masks blocks of time and frequency channels in the spectrogram, forcing the model to be robust to partial signal loss. Additionally, we used MixUp (Zhang et al., 2018), which trains the model on convex combinations of pairs of examples and their labels, smoothing the decision boundaries and improving generalization for rare classes.

3.4. Experimental Setup for Intersection Analysis

For the comparative analysis on the intersection dataset (90 species), we adopted a unified experimental protocol to ensure fair comparison between visual and acoustic models.

Data Splits The intersection dataset was partitioned using a stratified split strategy to preserve the class distribution across subsets. We used 70% of the data for training, 15% for validation, and 15% for testing. This differs slightly from the full CUB-200 splits to accommodate the smaller sample size of the intersection subset while ensuring sufficient validation data.

Training Configuration All models were trained for a maximum of 50 epochs with early stopping (patience of 10 epochs) based on validation loss. We used a batch size of 32 and Automatic Mixed Precision (AMP) to optimize training efficiency. Optimization strategies were tailored to each architecture:

- **Image ResNet-18:** Trained with SGD (learning rate 10^{-2}, momentum 0.9, weight decay 10^{-4}) and a StepLR scheduler (decay by 0.1 every 10 epochs).
- **Image ViT:** Trained with AdamW (learning rate 10^{-4}, weight decay 10^{-2}) and a Cosine Annealing scheduler ($T_{max} = 50$).
- **Audio CNN:** Trained with Adam (learning rate 10^{-3}).
- **Audio ViT:** Trained with Adam (learning rate 10^{-4}) and a Cosine Annealing scheduler.

4. Results

4.1. Image Classification

For each classification method developed in this study, we present below the corresponding evaluation metrics on the full CUB-200-2011 dataset (200 classes).

4.1.1. Traditional Methods

For the attribute-based traditional classifiers, the SVM achieved the best performance, obtaining an average accuracy of approximately 0.47 and a similar average F1-score under 5-fold cross-validation. This outcome was expected, given the SVM's ability to capture complex decision boundaries in high-dimensional spaces. Nevertheless, there remains substantial room for improvement, especially consid-

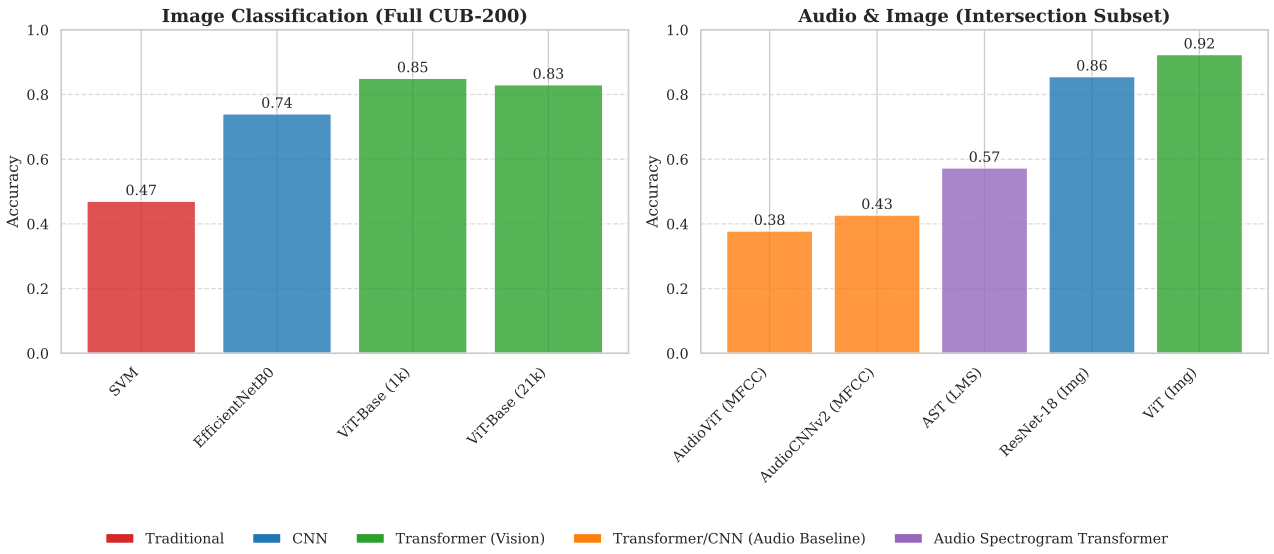

Figure 2. Comparative performance of visual and acoustic models. Left: Accuracy on the full CUB-200 dataset for image models. Right: Accuracy on the intersection dataset for audio and multimodal baselines. Vision Transformers (ViT) dominate both tasks, while AST significantly outperforms MFCC-based audio baselines.

ering the inherent limitations of manually curated attributes in capturing fine-grained visual variability.

Table 1. Traditional vs. Deep Learning Baselines (CUB-200)

Method	**Accuracy**	**F1 Macro**
SVM	0.471 ± 0.003	0.475 ± 0.002
LDA	0.467 ± 0.009	0.462 ± 0.007
Logistic	0.454 ± 0.012	0.460 ± 0.010
Random Forest	0.474 ± 0.003	0.453 ± 0.003
KNN (k=5)	0.300 ± 0.010	0.292 ± 0.008
Naive Bayes	0.188 ± 0.004	0.174 ± 0.006
Decision Tree	0.156 ± 0.002	0.143 ± 0.002
EfficientNetB0	0.74	0.74

4.1.2. Convolutional Neural Networks

In the CNN-based experiments on the full CUB-200 dataset, we evaluated EfficientNetB0. EfficientNetB0 demonstrated strong performance (74% accuracy). For EfficientNetB0, the base network was loaded without the fully connected head and using global average pooling. Fine-tuning was conducted by unfreezing only the top 25 layers, while the earlier convolutional layers remained frozen to preserve low-level pretrained representations. The classification head consisted of a Batch Normalization layer, followed by Dropout (0.25), a dense projection of 512 units with ReLU activation, an additional Dropout layer (0.25), and a final softmax output.

4.1.3. Vision Transformers

Finally, the Vision Transformer models achieved the strongest results among all evaluated methods. Both models outperformed the previous approaches. Surprisingly, the google/vit-base-patch16-224 model pretrained on ImageNet-1k consistently outperformed google/vit-base-patch16-224-in21k, despite the latter having been pretrained on a much larger and semantically richer dataset. This result contradicts standard literature benchmarks, which typically show benefits from larger pretraining datasets for fine-grained tasks (Kornblith et al., 2019). We hypothesize that the ImageNet-21k weights, being more generalized, required a longer warm-up phase or a smaller learning rate to adapt to the specific fine-grained features of CUB-200, whereas the ImageNet-1k weights were already sharper for object-centric classification.

Another unexpected finding was that advanced data-augmentation techniques (MixUp, CutMix, and RandAugment) did not improve performance, even though Vision Transformers are known to benefit from larger and more diverse training distributions. We observed that strong regularization (MixUp) harmed performance (85% $\rightarrow$ 81%). We attribute this to the high inter-class similarity in fine-grained tasks; blending images of nearly identical species likely confuses the model by smoothing decision boundaries that need to be extremely sharp. Still, the obtained results are acceptable for a fine-grained species-identification task, and indicate that further hyperparameter optimization could yield additional gains.

Table 2. Performance Comparison Across Augmentations. Note that the baseline (None) performed best.

Model	None		Mix+Cut		Mix+Cut+Rand	
	F1	Acc	F1	Acc	F1	Acc
ViT-Base (1k)	**0.85**	**0.85**	0.81	0.81	–	–
ViT-Base (21k)	0.83	0.83	0.80	0.80	0.83	0.83

Table 3. Summary of Best Models by Category

Category	Best Model	Accuracy	Dataset
Traditional	SVM	0.47	CUB-200
CNN	EfficientNetB0	0.74	CUB-200
Transformer	ViT-Base	0.92	Intersection

4.1.4. Image Classification on Intersection Dataset

To establish a fair baseline for our audiovisual experiments, we also evaluated our image models on the 90-species intersection dataset. We trained a ResNet-18 model specifically for this subset. While both models utilized ImageNet pretrained weights, the fine-tuning strategy differed slightly to accommodate the smaller dataset size. For ResNet-18, we unfroze the final fully connected layer and the last convolutional block, whereas the EfficientNetB0 approach involved unfreezing the top 25 layers. This model achieved an accuracy of 85.52%. On this same subset, the ViT-B/16 model achieved a test accuracy of 92.33%, significantly outperforming both the ResNet-18 baseline and the audio-only models. This high performance confirms that visual features remain the dominant modality for bird identification, though audio provides complementary information for specific cases.

4.2. Audio Classification Results

We present the progression of our audio classification models in Table 4.

Table 4. Audio Classification Performance (90-Species Intersection)

Phase / Model	Features	Strategy	Accuracy	F1 Macro
Phase 0: AudioViT	MFCC	Baseline	37.79%	0.13
Phase 2: AudioCNNv2	MFCC	Focal Loss	42.72%	0.22
Phase 3: AST	LMS	Transfer Learning	**57.28%**	**0.36**

Analysis Our baseline AudioViT using MFCCs achieved only 37.79% accuracy. Notably, this experiment demonstrated that direct transfer learning from ImageNet-pretrained Vision Transformers to MFCC tensors failed to prove useful, performing worse than a simple Convolutional Neural Network (AudioCNNv2) which achieved 42.72%. Examination of the training dynamics revealed severe and early overfitting: while training accuracy surged to nearly 99% within just 10 epochs, validation loss began to diverge and increase as early as epoch 4. This indicates that the model simply memorized the training set noise rather than learning generalizable spectro-temporal features. This negative result highlights the limitations of cepstral coefficients for deep learning models that thrive on spatial correlations and suggests that the domain gap between natural images and MFCC heatmaps is too large for effective transfer without more suitable input representations. The introduction of Focal Loss in Phase 2 (AudioCNNv2) further improved performance, demonstrating the importance of addressing the severe class imbalance (608:1).

The most significant leap occurred in Phase 3 with the adoption of the Audio Spectrogram Transformer (AST) and Log-Mel Spectrograms (LMS). This combination achieved 57.28% accuracy, a relative improvement of nearly 35% over the baseline. This confirms that (1) LMS provides a superior representation for vision-based architectures applied to audio, and (2) the attention mechanism of AST effectively captures the temporal dynamics of bird calls better than fixed-window CNNs.

4.3. Future Work: Multimodal Fusion

Given the strong performance of our separate modalities, this work establishes a solid foundation for future multimodal integration. Our comparative analysis demonstrates that while vision is dominant, audio provides a distinct and complementary signal. Future work will focus on an *Intermediate Feature Fusion* strategy. Rather than simple late fusion (averaging predictions), we aim to concatenate the embedding vectors from the penultimate layers of the fine-tuned ViT and AST models. These concatenated vectors will be fed into a joint Multi-Layer Perceptron (MLP) (Rosenblatt, 1958) to learn correlations between visual features (plumage, shape) and acoustic features (call structure), potentially resolving cases where one modality is ambiguous.

5. Discussion

Our results highlight the efficacy of Transformer-based architectures for fine-grained bird classification in both visual and acoustic modalities. The Vision Transformer (ViT) achieved the highest overall accuracy (92.33% on the intersection dataset), outperforming the ResNet-18 baseline (85.52%). This suggests that the self-attention mechanism in ViT is particularly well-suited for capturing the subtle, non-local discriminative features (such as plumage patterns and beak shapes) that distinguish similar bird species, even with a relatively small dataset like CUB-200. The strong performance of ViT, despite the lack of inductive bias typical of CNNs, can be attributed to the effective transfer learning from the large-scale ImageNet-21k pretraining.

In the audio domain, the transition from MFCCs to Log-Mel Spectrograms (LMS) proved critical. MFCCs, while standard for speech, discard too much spectro-temporal information, limiting the performance of deep models (37.79% accuracy). By treating audio as an image via LMS, we enabled the use of powerful vision architectures. The Audio Spectrogram Transformer (AST) further capitalized on this by achieving 57.28% accuracy, a massive improvement over the CNN baselines. This indicates that bird calls, which often have complex temporal structures and long-range dependencies (e.g., repeated motifs), benefit significantly from the global receptive field of Transformers.

We also observed that addressing class imbalance was essential. The use of Focal Loss in Phase 2 provided a clear performance boost, validating its utility in biodiversity datasets where species distributions are naturally long-tailed.

Training Dynamics and Generalization A critical analysis of the training curves revealed distinct behaviors between modalities (Table 5). The image-based Vision Transformer exhibited robust generalization, with training accuracy reaching 97% and test accuracy maintaining 92%, indicating that the large-scale pretraining on ImageNet-21k effectively bridged the domain gap. In contrast, all audio models suffered from significant overfitting due to the smaller size of the intersection dataset. The AudioViT baseline memorized noise early (epoch 4), while the AudioCNNv2 showed optimization instability with diverging validation loss after epoch 10. The AST model, despite achieving the best audio performance, also displayed a disconnect between loss and accuracy: validation loss began to rise after epoch 3, even as accuracy continued to improve until epoch 10. Similar overfitting on noisy, crowd-sourced audio data has been observed in other fine-grained audiovisual studies (Horn et al., 2022). This suggests that while Transformers are powerful, they require careful regularization or larger datasets to fully mitigate overfitting in the acoustic domain.

Table 5. Training Dynamics Analysis: Comparison of peak training accuracy, validation accuracy, and the epoch where validation loss began to diverge (overfitting onset).

Model	Modality	Train Acc	Val Acc	Overfitting Onset
ImageViT	Visual	97%	92%	None (Stable)
AudioViT	Audio	99%	38%	Epoch 4
AudioCNNv2	Audio	95%	43%	Epoch 10
AST	Audio	98%	57%	Epoch 3

6. Conclusion

This work presented a comprehensive evaluation of deep learning methodologies for fine-grained audiovisual categorization of birds. We demonstrated that: 1. **Transformers dominate:** ViT and AST consistently outperformed their convolutional counterparts in both image and audio tasks. 2. **Representation matters:** Log-Mel Spectrograms are superior to MFCCs for deep learning-based audio classification. 3. **Imbalance handling is key:** Techniques like Focal Loss are necessary to handle real-world biodiversity data distributions.

While image classification remains the more accurate modality (92.33%), our audio models showed significant promise (57.28%), suggesting that acoustic data contains complementary information. Future work will focus on the proposed multimodal fusion strategy to leverage the strengths of both domains, aiming for a robust, holistic bird identification system.

Acknowledgements

If a paper is accepted, the final camera-ready version can (and usually should) include acknowledgements. Such acknowledgements should be placed at the end of the section, in an unnumbered section that does not count towards the paper page limit. Typically, this will include thanks to reviewers who gave useful comments, to colleagues who contributed to the ideas, and to funding agencies and corporate sponsors that provided financial support.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here."

References

Abdel-Hamid, O., Mohamed, A.-r., Jiang, H., Deng, L., Penn, G., and Yu, D. Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(10):1533–1545, 2014.

Anusha, P. and ManiSai, K. Bird species classification using deep learning. In *2022 International Conference on Intelligent Controller and Computing for Smart Power (ICICCSP)*, pp. 1–5, 2022. doi: 10.1109/ICICCSP53532.2022.9862344.

Breiman, L. Random forests. *Machine Learning*, 45(1): 5–32, 2001.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. *Classification and regression trees*. CRC press, 1984.

Cortes, C. and Vapnik, V. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

Cover, T. and Hart, P. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

Cox, D. R. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2):215–232, 1958.

Cubuk, E. D., Zoph, B., Shlens, J., and Le, Q. V. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 702–703, 2020.

Das, S. D. and Kumar, A. Bird species classification using transfer learning with multistage training, 2018. URL https://arxiv.org/abs/1810.04250.

Davis, S. and Mermelstein, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366, 1980.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

Fisher, R. A. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 1936.

Ghani, B. et al. Global bird species recognition in real-time. *arXiv preprint arXiv:2307.16384*, 2023.

Gong, Y., Chung, Y.-A., and Glass, J. AST: Audio Spectrogram Transformer. In *Proc. Interspeech 2021*, pp. 571–575, 2021. doi: 10.21437/Interspeech.2021-698.

Hand, D. J. and Yu, K. Idiot's bayes-not so stupid after all? *International Statistical Review*, 69(3):385–398, 2001.

He, J., Chen, J.-N., Liu, S., Kortylewski, A., Yang, C., Bai, Y., and Wang, C. Transfg: A transformer architecture for fine-grained recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 852–860, 2022.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Horn, G. V., Qian, R., Wilber, K., Adam, H., Aodha, O. M., and Belongie, S. Exploring fine-grained audiovisual categorization with the ssw60 dataset, 2022. URL https://arxiv.org/abs/2207.10664.

Islam, S., Khan, S. I. A., Abedin, M. M., Habibullah, K. M., and Das, A. K. Bird species classification from an image using vgg-16 network. In *Proceedings of the 7th International Conference on Computer and Communications Management*, ICCCM '19, pp. 38–42, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450371957. doi: 10.1145/3348445.3348480. URL https://doi.org/10.1145/3348445.3348480.

Kahl, S., Wood, C. M., Eibl, M., and Klinck, H. Birdnet: A deep learning solution for avian diversity monitoring. *Ecological Informatics*, 61:101236, 2021.

Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., and Plumbley, M. D. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020.

Kornblith, S., Shlens, J., and Le, Q. V. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2661–2671, 2019.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2980–2988, 2017.

Marini, A., Facon, J., and Koerich, A. L. Bird species classification based on color features. In *2013 IEEE International Conference on Systems, Man, and Cybernetics*, pp. 4336–4341, 2013. doi: 10.1109/SMC.2013.740.

Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., and Le, Q. V. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In *Proc. Interspeech 2019*, pp. 2613–2617, 2019. doi: 10.21437/Interspeech.2019-2680.

Rosenblatt, F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

Tan, M. and Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pp. 6105–6114. PMLR, 2019.

Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Xeno-canto Foundation. Xeno-canto: Sharing wildlife sounds from around the world, 2025. URL https://xeno-canto.org/.

Yang, Z., Luo, T., Wang, D., Hu, Z., Gao, J., and Wang, L. Learning to navigate for fine-grained classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 420–435, 2018.

Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., and Yoo, Y. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6023–6032, 2019.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=r1Ddp1-Rb.