**Efficient Fine-Grained Bird Classification via Dynamic Ensembles and Active Learning**

# Abstract: Fine-grained bird species classification is a challenging task, mainly because the differences between species are often very subtle, and there is typically limited labeled data available for each species. This paper introduces a novel approach that combines multiple deep-learning models with a smart labeling strategy to address these issues. The method uses a dynamic ensemble of three advanced models: ResNet50, which excels in recognizing visual patterns; EfficientNet-B3, known for its efficiency and performance balance; and Swin Transformer, a newer model that captures both local and global features in images. Instead of relying on each model’s prediction separately, an adaptive mechanism (a multi-layer perceptron or MLP) weighs their predictions per sample, allowing the system to make smarter and more accurate decisions. Overall, the approach stands out due to its combination of a dynamic ensemble of models and an active learning strategy. This not only boosts performance but also makes the best use of limited labeled data, offering a more efficient way to tackle complex classification tasks like bird species identification. The results demonstrate that with fewer labels, the method can still achieve high accuracy and generalize well, providing a powerful tool for fine-grained classification in scenarios where data is scarce.

# Keywords: Fine-grained, Classification, Ensemble, ResNet50, EfficientNet-B3, Swin Transformer, MLP, Active Learning, Limited Data, Accuracy

# INTRODUCTION

#  Distinguishing between bird species is a particularly challenging task in computer vision because many species appear extremely similar, often differing only in subtle features like feather patterns or beak shape. This challenge is further amplified by the limited amount of labeled training data typically available for each category. A well-known example of such complexity is the CUB-200-2011 dataset [1], which includes 200 bird species that can be hard to tell apart even for humans. While deep learning models like ResNet and newer architectures have significantly improved image classification in general, they often struggle with fine-grained recognition due to their difficulty in capturing minute differences. Additionally, annotating large-scale datasets with hundreds of categories is a time-consuming and resource-intensive process. To address these issues, ensemble learning, where multiple models are combined to improve performance has proven effective, as it not only enhances accuracy but also provides more reliable uncertainty estimates. These estimates are especially useful in active learning, a method that reduces the need for extensive labeled data by selectively choosing the most informative samples to annotate.

# However, the combination of ensemble methods and active learning has not been widely explored for fine-grained classification. In this work, we introduce a dynamic ensemble system that integrates active learning to tackle this gap. Our method combines the strengths of three advanced models ResNet50, EfficientNet-B3, and Swin Transformer treating them as expert classifiers. A small neural network learns to adaptively weight its predictions for each image, effectively deciding which model to trust more in different scenarios. Alongside this, we use an active learning strategy that identifies and prioritizes the most uncertain, and therefore informative, images for labeling. This targeted approach helps the model learn more efficiently from fewer labeled examples. Our experiments on the CUB-200-2011 dataset show that this combination not only boosts classification performance but also reduces the need for large amounts of annotated data, outperforming individual models and static ensembles in multiple accuracy and performance metrics.

# RELATED WORK

Fine-grained image classification, such as recognizing subtle differences between bird species, remains a complex challenge in computer vision. Many bird classes look nearly identical, and distinguishing them often requires focusing on tiny visual cues like feather patterns or beak shapes. Early research in this area attempted to improve accuracy by detecting and aligning specific object parts, such as the head or wings. Later, more advanced models like bilinear CNNs combined information from two separate neural networks to better capture detailed textures [2]. More recent approaches use attention mechanisms and transformer-based models to understand both fine details and overall image context, and some even integrate extra data like labeled attributes. However, despite these advances, fine-grained tasks are still difficult, especially when there's not enough labeled data or when the dataset is imbalanced, with some classes underrepresented [3].

One widely used technique to address these challenges is ensemble learning, which combines multiple models to create a more robust and accurate system. Traditional ensemble methods like bagging and boosting have long been known to improve performance by reducing errors and balancing model biases. In the deep learning world, ensembles are especially popular in high-stakes applications, as they often outperform single models and provide more stable predictions. Typically, ensembles are static, meaning they average or vote on the outputs of several fixed models. However, newer research is exploring dynamic ensembles, where the way model outputs are combined changes depending on the input [4]. This approach, inspired by the mixture-of-experts concept, uses a small neural network to learn how to assign different importance to each model for each input. For instance, a model that’s especially good at recognizing certain visual traits might be given more influence when those traits are present. In our work, we use this idea to combine the strengths of convolutional models like ResNet50 and EfficientNet with transformer-based models like Swin Transformer, resulting in a smarter, input-sensitive ensemble system [5].

Alongside this, we integrate active learning, which helps reduce the amount of data that needs to be labeled. Instead of randomly choosing which images to annotate, the model selects the ones it’s most uncertain about—those that are likely to teach it the most [6]. This process ensures that each new label has a high impact on improving performance. Several strategies exist for this, such as picking the most uncertain predictions, selecting samples where models disagree, or ensuring a diverse sample pool. Deep learning researchers have also developed methods to estimate model uncertainty more effectively, either by using techniques like dropout during inference or by relying on ensembles, which tend to produce more reliable uncertainty estimates than individual models [7]. However, most of these methods assume that the ensemble models are trained independently and are not integrated tightly.

# METHODOLOGY

## Dataset and Preprocessing: Our experiments are conducted using the CUB-200-2011 bird dataset, a widely used benchmark for fine-grained image classification. This dataset includes 11,788 images across 200 different bird species, with approximately 30 images per species reserved for training (5,994 images total) and a comparable number for testing (5,794 images). The images are high-quality photographs taken in natural environments, featuring a wide range of variations in lighting, poses, and backgrounds. These conditions, along with the visual similarities between many bird species, make the dataset particularly well-suited for evaluating fine-grained recognition techniques. In our study, we focus solely on image classification and do not utilize the dataset’s additional annotations, such as part locations or attributes. To prepare the images for model training, all input images are resized to a uniform size—typically 224×224 pixels—to meet the input requirements of the models we use. To improve the generalization of our models and make them more robust to variations in image conditions, we apply several standard data augmentation techniques during training. These include random horizontal flipping, cropping (by selecting a random 224×224 patch from a resized image), and light color jittering. These augmentations help simulate real-world variability and reduce overfitting. Additionally, pixel values are normalized using mean and standard deviation values from the ImageNet dataset, as our base models are pre-trained on ImageNet. During testing, no augmentations are applied except for center cropping, ensuring a consistent and fair evaluation.

## Base Models and Dynamic Ensemble: Our approach brings together the strengths of three advanced image recognition models—ResNet50, EfficientNet-B3, and Swin Transformer—to tackle the challenging task of fine-grained bird species classification. Each model is chosen for its unique capabilities: ResNet50, a well-established convolutional neural network, is particularly strong at extracting local patterns and textures thanks to its use of residual connections. EfficientNet-B3 is a more recent convolutional model that achieves an excellent balance between accuracy and efficiency by smartly scaling its depth, width, and resolution. The Swin Transformer, based on the transformer architecture, processes images using a series of shifted windows, enabling it to capture both small details and the broader structure of an image [8]. By using these three models together, we cover a wide range of visual features—fine textures, structural patterns, and global context.

We start by fine-tuning all three models on the CUB-200-2011 bird dataset, using pre-trained weights from ImageNet as a starting point. This speeds up training and improves performance, especially since the dataset has relatively few images per class. Instead of treating each model’s predictions equally, we build a more intelligent system that adapts to the content of each image. This is achieved through a small neural network—known as a gating network—that learns to decide how much to rely on each model for a given input. For example, if a bird image has strong texture cues, the system might rely more on EfficientNet. If the overall shape or posture is more informative, it may favor the Swin Transformer. The gating [9] network works by analyzing the outputs from all three base models and assigning them adaptive weights. These weights determine how much each model contributes to the final prediction, allowing the system to dynamically adjust based on the specific features of each image. This approach goes beyond traditional ensembles that use fixed rules, like averaging or voting, by making model combinations a learned and flexible process.

We train the entire ensemble, including the base models and the gating network, together as one unified system. This allows all components to learn in coordination, optimizing not just individual model performance, but how they work together. By letting the system decide which model to trust more for each image, we achieve more accurate and nuanced classifications, especially in cases where bird species are visually very similar. An architectural overview (illustrated in Figure 1) shows how the base models feed into the gating network, which then combines their predictions into a final decision.

## Active Learning Loop: Labeling thousands of bird images can be both time-consuming and expensive, especially since it often requires the expertise of ornithologists or trained annotators to accurately identify species that may look nearly identical. To reduce this burden while still maintaining high model performance, we incorporate an active learning loop into our training process. This strategy is designed to help the model learn efficiently by selectively choosing the most informative images to label, rather than labeling the entire dataset up front [10].

We rely on a technique called entropy-based uncertainty sampling to guide this process. Entropy, in this context, measures how uncertain the model is about a particular prediction. If the model assigns nearly equal probabilities to many different classes for a single image, it means it is unsure, which makes that image a valuable candidate for labeling. The more uncertain the prediction, the more the model can potentially learn from that sample once the correct label is known.

Here’s how the active learning loop works: we begin with a small, randomly selected subset of labeled training data (about 10% of the full training set), and use it to train our dynamic ensemble model. The remaining training images are left unlabeled and form the unlabeled pool. Once the model is trained on the initial labeled set, we use it to predict labels for every image in the unlabeled pool and calculate the uncertainty (entropy) of each prediction. The top 50 most uncertain images are then selected for labeling in that round. These selected images are sent to an oracle—typically a human expert or annotator—for labeling. In our experiments, we simulate this step using the known labels from the dataset. Once labeled, these new examples are added to the training set, and the process repeats: the model is retrained (or fine-tuned) using the expanded labeled set, and the next batch of uncertain images is selected. We continue this cycle for a set number of rounds or until the labeling budget is used up.

This strategy allows the model to focus on the most ambiguous and challenging images, like visually similar species of sparrows, early in the process, which helps it improve more quickly. As more informative samples are added, the model becomes increasingly confident, and the average uncertainty of the unlabeled pool drops. We also speed up training by carrying over the model weights from the previous round, avoiding the need to retrain from scratch each time.

## Training Protocol: To effectively train our models, we use the Adam optimizer, which is well-suited for deep learning tasks due to its adaptive learning rate capabilities. We set the optimizer's parameters to commonly used values that work well in practice and start with a small learning rate to allow fine-tuning of both the base models and the gating network with stability. During each training phase—whether in the initial full-data scenario or throughout the active learning rounds—we train the model for up to 50 epochs using mini-batches of 32 images. To avoid overfitting and ensure efficient training, we monitor the model’s performance on a small validation set (10% of the training data, distinct from the test set). If the model’s accuracy on the validation set doesn’t improve for five consecutive epochs, training is stopped early.

When training the dynamic ensemble, we use a carefully designed two-phase approach. Initially, each of the three base models—ResNet50, EfficientNet-B3, and Swin Transformer—is trained individually for a few epochs to ensure each model develops a reasonable understanding of the data. Only after this initial phase do we introduce the gating network, which learns to combine its outputs. This staged training approach helps avoid situations where the gating model prematurely favors one base model before the others have learned meaningful features. To further encourage balanced learning, we apply a small penalty (L2 regularization) to the gating network to prevent it from becoming overly biased toward any single model. As new labeled data is added during active learning rounds, all components—the base models and the gating MLP—continue to be fine-tuned using the expanded dataset.

Our implementation is built using PyTorch. For initialization, we leverage pre-trained weights: ResNet50 and EfficientNet-B3 use models from the torch vision library, while Swin Transformer is initialized from a publicly released checkpoint. The gating MLP itself is a simple neural network with one hidden layer of 256 neurons and ReLU activation. To further refine training, we reduce the learning rate whenever the validation loss stops improving for a few epochs.

# EXPERIMENTS AND RESULTS

## Evaluation Metrics: We also use macro-precision, macro-recall, and macro-F1 scores to get a more balanced view of how the model is doing across all bird species. These metrics are calculated for each species individually and then averaged, so every bird class is treated equally, no matter how difficult or easy it is to recognize. This matters because the dataset is balanced, with about the same number of images per species, so we want to ensure the model performs well across the board, not just on the more distinctive birds. To dig deeper into the model’s behavior, we examine its confusion matrix, which shows which species are often mistaken for others. This helps us see patterns, like whether the model regularly confuses two nearly identical birds, and gives insight into where the model might still struggle. All results are measured using a separate test set that the model never saw during training or the active learning process. This helps ensure that the results truly reflect how the model would perform on new, unseen data.

## Baselines: To evaluate the effectiveness of our approach, we compare it against a range of baseline methods that help highlight the individual contributions of dynamic ensembling and active learning. First, we include three strong single-model baselines: ResNet50, a classic convolutional neural network; EfficientNet-B3, a compact yet high-performing CNN; and the Swin Transformer, which represents the growing use of transformer architectures in vision tasks. All of these are fine-tuned on the full training dataset using standard supervised learning. We then evaluate a static ensemble that simply averages the predictions of these three models using equal weights—this helps us understand the benefit of ensembling without any adaptive weighting. Next, we introduce a dynamic ensemble model, which uses a gating MLP to intelligently weigh the outputs of the base models. This version is trained on the full labeled dataset and helps isolate the impact of dynamic weighting alone, without involving active learning. Finally, we assess our full method, which combines dynamic ensembling with an active learning strategy that selects the most informative samples for labeling. We report results both under limited labeling budgets (e.g. when only 30% of the data is labeled through active learning) and when the active learning process continues until all data is labeled. All methods follow the same training and validation protocols to ensure fairness, except active learning, which gradually expands the labeled dataset during training. It's worth noting that the static ensemble uses fixed, equal weights without tuning, meaning that with optimized weights, it could potentially perform slightly better, but our goal here is to compare reasonable baselines rather than exhaustively fine-tune them.

## Top of Form

## Bottom of Form

**4.3 Final Results:** We gathered all of our numbers into one place to see how each method stacks up on the CUB-200-2011 test set. As you’d expect, the three standalone networks each do pretty well on their own: ResNet50 lands around 84.7% Top-1 accuracy, EfficientNet-B3 edges up to 86.5%, and the Swin Transformer nudges a bit higher at 87.2%. When we simply average their predictions in a static ensemble, that small boost adds up – we jump to 89.0% Top-1, which confirms that combining diverse models is a quick way to shave off a few more mistakes.

But the real star is the dynamic ensemble, where a tiny gating network decides on-the-fly how much to trust each expert for each bird photo. By letting those weights vary per image rather than holding them fixed, we push Top-1 accuracy to 90.5%, and Top-3 to 97.8%. We also see macro-precision climb to 0.91 and macro-F1 to about 0.905, meaning fewer false alarms and a more balanced performance across all 200 species.

Finally, we peeked into the confusion matrix to find the stubborn mistakes. Most birds are recognized reliably, but nearly identical pairs—like Western Grebe versus Clark’s Grebe—still trip us up, since distinguishing them visually is a real challenge even for experts. On the bright side, the ensemble does a much better job than any single model at teasing apart look-alike sparrows and other tricky groups. That tells us our gating network is learning to focus on the subtle color or shape cues that one model alone might overlook.

Table 1: Final performance on CUB-200-2011
Top-1 and Top-3 accuracy, plus macro-averaged precision, recall, and F1. The Active Learning (AL) result uses 50% labeled data; all other methods use 100% labeled data.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Method**  | **Top-1**  | **Top-3**  | **Macro-P**  | **Macro-R**  | **Macro-F1**  |
| ResNet50 (single) [2]  | 84.7%  | 95.0%  | 0.85  | 0.84  | 0.84  |
| EfficientNet-B3 (single) [3]  | 86.5%  | 95.8%  | 0.87  | 0.86  | 0.86  |
| Swin Transformer (single) [4]  | 87.2%  | 96.0%  | 0.87  | 0.87  | 0.87  |
| Static Ensemble (3-model avg)  | 89.0%  | 97.0%  | 0.89  | 0.89  | 0.89  |
| **Dynamic Ensemble (ours)**  | **90.5%**  | **97.8%**  | **0.91**  | **0.90**  | **0.905**  |
| Dynamic Ens. + AL (50% labels)  | 87.0%  | 95.5%  | 0.86  | 0.85  | 0.86  |

# DISCUSSION

Our experiments clearly show that both parts of our system—dynamic weighting and active learning—matter a great deal for fine-grained bird classification.

First, the dynamic ensemble consistently beats a static one. Giving each model a fixed share of the vote simply doesn’t cut it when you have hundreds of nearly identical species. By contrast, our little gating network can spot when one of the experts is struggling (say, because the photo is grainy or the bird is obscured) and shift trust toward the more reliable model at that moment. Likewise, if the image is crisp and full of detail, it might lean on the Transformer’s global view. This on-the-fly flexibility is what drives our bump in accuracy and precision.

Second, active learning delivers huge efficiency gains. Rather than labeling thousands of random bird photos, we let the ensemble tell us which images it finds most confusing, and we only ask experts to label those. Amazingly, with only half the labels, we nearly match the performance of a model trained on every single annotation. In real-world settings—ecological surveys, wildlife monitoring, or citizen-science projects—this could save weeks or months of expert time, while still giving you a top-tier classifier. Plus, active learning naturally focuses on the toughest species, which helps keep performance balanced across all classes (reflected in our strong macro-metrics).

That said, our solution isn’t perfect. Running three hefty networks plus a gating MLP for every image can be slow, so it might not suit real-time applications without further work. Model compression or distilling the ensemble into a single, leaner network could help there. Although active learning cuts labeling costs, it still needs an initial seed set and someone (an “oracle”) to provide ground-truth labels. In fields where getting even a small number of labels is hard, blending in semi-supervised or self-supervised techniques could push efficiency even further.

Looking ahead, there is plenty of room to grow. Our current gate decides based on each model’s confidence scores, but more sophisticated strategies—perhaps using image metadata, geolocation, or part-detection cues—might help when all experts are equally stumped. While birds were our focus, this mix of dynamic ensembling and smart sampling could easily transfer to other fine-grained tasks—identifying car models, plant species, or even medical images.

In short, by marrying an adaptable “who to trust” mechanism with focused, uncertainty-driven labeling, we bridge the gap between raw performance and practical efficiency. We believe this approach can pave the way for deployable, high-accuracy recognition systems that make the most of expert effort and diverse model strengths.

# CONCLUSION

# We presented a dynamic ensemble and active-learning framework for fine-grained bird species classification on the challenging CUB-200-2011 dataset. The dynamic ensemble combines a CNN backbone (ResNet50), an efficient CNN (EfficientNet-B3), and a vision Transformer (Swin) with an adaptive weighting scheme, yielding superior accuracy over traditional ensembles or single models. Coupling this with an entropy-based active learning loop allowed us to reach competitive performance with far fewer labeled samples, addressing the high cost of annotations in fine-grained tasks. Our best model achieved a Top-1 accuracy of over 90% on 200 bird species, and even with 50% of the training data labeled, it surpassed baseline models that used all data.

This work contributes to state-of-the-art fine-grained visual recognition by demonstrating that model heterogeneity and data-centric learning strategies (active sample selection) can be jointly leveraged. In terms of broader impact, such techniques could facilitate the creation of accurate biodiversity monitoring tools or field guides that require minimal expert annotation, thereby aiding conservation efforts. Future work will explore compressing the dynamic ensemble for real-time use and extending the active learning strategy with semi-supervised learning to further reduce labeling requirements. We also plan to apply our method to other fine-grained datasets (e.g., insect species, and plant diseases) to validate its generality. Ultimately, our approach highlights the promise of combining *who* (an ensemble of “experts”) and *what* (active querying of data) to efficiently tackle difficult recognition problems.

#

# REFERENCES

1. C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, **“The Caltech-UCSD Birds-200-2011Dataset,”** California Institute of Technology, Tech. Report CNS-TR-2011-001, 2011.
2. K. He, X. Zhang, S. Ren, and J. Sun, **“Deep Residual Learning for Image Recognition,”** in *Proc. IEEE CVPR*, 2016, pp.770–778.
3. M. Tan and Q.V. Le, **“EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,”** in *Proc. ICML*, vol. 97, 2019, pp. 6105–6114.
4. Z. Liu *et al*., **“SwinTransformer: Hierarchical Vision Transformer Using Shifted Windows,”** in *Proc. IEEE ICCV*, 2021, pp. 9992–10002.
5. T.-Y. Lin, A., Roy Chowdhury, and S. Maji, **“Bilinear CNN Models for Fine-Grained Visual Recognition,”** in *Proc. IEEE ICCV*, 2015, pp. 1449–1457.
6. T.G. Dietterich, **“Ensemble Methods in Machine Learning,”** in *Proc. MCS (LNCS1857)*, 2000, pp. 1–15.
7. B. Settles, **“Active Learning Literature Survey,”** University of Wisconsin–Madison, Computer Sciences Technical Report, 2010.
8. Y. Gal, R. Islam, and Z. Ghahramani, **“Deep Bayesian Active Learning with Image Data,”** in *Proc. ICML*, 2017, pp.1183–1192.
9. W.H. Beluch, T. Genewein, A. Nürnberger, and J.M. Köhler, **“The Power of Ensembles for Active Learning in Image Classification,”** in *Proc. IEEE CVPR*, 2018, pp. 9368–9377.
10. O. Sener and S. Savarese, **“Active Learning for Convolutional Neural Networks: A Core-Set Approach,”** in *Proc. ICLR*, 2018.