Fine-Grained Bird Species Classification on CUB-200 with ResNet-50

Shiheng Xu Renjie Fan Kunshu Yang

Abstract

Fine-grained visual classification (FGVC) of bird species demands sensitivity to subtle inter-class variations and robustness against real-world noise. In this work, we present a single-model ResNet-50 pipeline on the official CUB-200-2011 train/test split (5,994/5,794 images) that systematically balances accuracy and efficiency. Building on the torchvision ResNet-50 implementation and open-source CBAM and MixUp repositories, we employ aggressive augmentations (RandomResizedCrop, ColorJitter, RandomErasing, MixUp), label smoothing, and weight decay; a phased fine-tuning schedule coupled with OneCycleLR accelerates convergence; and CBAM attention modules sharpen part-focused representations. Our approach achieves $79.2\% \pm 0.1\%$ Top-1 ($95.1\% \pm 0.1\%$ Top-5) on the held-out test set, with validation accuracy stabilizing above 77% by epoch 20 and only minimal overfitting. A targeted error-mode analysis guides next-step augmentations for lighting and occlusion challenges.

Introduction

Fine-grained visual classification (FGVC) of bird species demands both sensitivity to very subtle inter-class differences and robustness against high intra-class variability under real-world conditions. In our early experiments, multi-stage and ensemble methods improved accuracy but introduced substantial complexity and long training times. Consequently, we refined our problem statement to focus on *designing an efficient, single-model pipeline based on ResNet-50* that delivers competitive accuracy while remaining practical for deployment.

We evaluate on the standard Caltech–UCSD Birds-200-2011 (CUB-200) benchmark [12], comprising 11,788 images across 200 species. To ensure direct comparability with prior work, we use the official train/test split (5,994 training / 5,794 test), without creating additional validation partitions. This setup highlights how targeted augmentations and attention can compensate for a lighter pipeline.

Our implementation builds on PyTorch's torchvision ResNet-50 repository and incorporates open-source CBAM attention modules [8] and MixUp code from Facebook Research [3]. We apply aggressive data augmentations (RandomResizedCrop, ColorJitter, RandomErasing, MixUp), label smoothing, and weight decay to regularize the network. A phased fine-tuning strategy—training only the new classification head for one epoch, then unfreezing the entire network—works in concert with a OneCycleLR schedule to accelerate convergence to a well-generalized minimum. Finally, CBAM modules sharpen the model's focus on discriminative bird parts, such as beaks and wing patterns.

Related Work

Deep residual networks such as ResNet-50 [11], provided via PyTorch's torchvision library, have become the de facto backbone for fine-grained visual classification (FGVC) thanks to their strong representational power. However, directly fine-tuning these models on small FGVC datasets often leads to overfitting.

To mitigate this, a variety of data augmentation and regularization techniques have been proposed. MixUp [3] (as implemented in Facebook Research's `mixup-cifar10` repository), CutMix [4], and RandomErasing [5] all synthetically enrich training distributions, while label smoothing [15] and weight decay further curb overconfidence. We incorporate these methods in our pipeline to improve robustness.

Learning-rate schedules play a critical role in convergence and generalization. The OneCycleLR policy [7], which ramps the learning rate up and then decays it within a single training cycle, has been shown to accelerate training and yield better minima than static schedules. We leverage PyTorch's `OneCycleLR` scheduler to implement this strategy.

Attention modules have proven effective at highlighting discriminative regions for FGVC. In particular, the Convolutional Block Attention Module (CBAM) [8], available from the jongchan/attention-module GitHub repository, adaptively refines both channel and spatial features. By integrating CBAM into ResNet-50, our model learns to focus on bird parts such as beaks, wing bars, and crowns.

Classic FGVC approaches—like RA-CNN [9] and MA-CNN [10]—rely on region proposals or part-based architectures to capture local details. In contrast, our work demonstrates that a single, end-to-end ResNet-50 pipeline, augmented with strong data transforms and lightweight attention, can achieve nearly 80% Top-1 accuracy on CUB-200 while maintaining simplicity and efficiency.

Methodology

Fully-Connected Classification Head

$$\hat{\mathbf{z}} = W\,\mathrm{GAP}\big(f_{\mathrm{conv}}(\mathbf{x})\big) + \mathbf{b}, \tag{1}$$

We adapt the ImageNet-pretrained ResNet-50 backbone [11] by replacing its 1000-way classifier with a 200-way linear projection. Global-average-pooled features $\mathbf{h}$ ($d = 2048$) are mapped to class logits $\hat{\mathbf{z}} \in \mathbb{R}^{200}$ by weights W and bias $\mathbf{b}$.

Phased Fine-Tuning Schedule

$$\eta_e = \begin{cases} 1 \times 10^{-3}, & e = 1, \\ 3 \times 10^{-4}, & e \in \{5, \ldots, 30\}, \end{cases} \tag{2}$$

Epoch1 trains only the new head at a high learning rate, then all layers are unfrozen from epoch5 onward. This staged strategy preserves pretrained features while allowing task-specific refinement.

Label-Smoothed Cross-Entropy Loss

$$\mathcal{L} = -\sum_{i=1}^{200} \tilde{y}_i \log\big((\hat{\mathbf{z}})_i\big), \qquad \tilde{y}_i = (1-\varepsilon)\mathbf{1}_{\{i=y\}} + \tfrac{\varepsilon}{200}, \tag{3}$$

We follow the label-smoothing formulation of Szegedy *et al.* [15] with $\varepsilon = 0.1$ to reduce over-confidence and improve generalisation.

AdamW Optimisation

$$\theta \leftarrow \theta - \eta \, \frac{m_t}{\sqrt{v_t} + \delta} - \lambda \, \theta, \tag{4}$$

Parameters are updated with AdamW [16]; $\lambda = 1 \times 10^{-4}$ decouples weight decay from the gradient step and accelerates convergence.

OneCycle Learning-Rate Policy

$$\eta_t = \begin{cases} \eta_{\max} \dfrac{t}{t_{\text{warm}}}, & t \le t_{\text{warm}}, \\ \eta_{\max} \dfrac{1 + \cos\!\big(\pi \frac{t - t_{\text{warm}}}{T - t_{\text{warm}}}\big)}{2}, & t > t_{\text{warm}}, \end{cases} \tag{5}$$

The OneCycle schedule of Smith & Topin [17] warms to $\eta_{\max} = 3 \times 10^{-4}$ over the first 10% of iterations, then cosine-anneals to near-zero for stable convergence.

Gradient Clipping

$$\nabla_\theta \mathcal{L} \leftarrow \nabla_\theta \mathcal{L} \cdot \min\!\Big(1, \frac{G_{\max}}{\|\nabla_\theta \mathcal{L}\|_2}\Big), \qquad G_{\max} = 5.0. \tag{6}$$

We cap gradient norms as recommended by Pascanu *et al.* [18] to prevent rare exploding updates.

These five components collectively raise single-model test accuracy to **77.46 %** on the CUB-200 dataset.

Datasets

CUB-200-2011 Bird Dataset

The Caltech–UCSD Birds-200-2011 benchmark (CUB-200) [12] is a standard testbed for *fine-grained* visual categorisation. It contains 11 788 colour photographs of **200** North-American bird species. Each image is annotated with (i) a species label, (ii) a tight bounding box, and (iii) fifteen part landmarks (beak, crown, wings, tail, *etc.*). The dataset's subtle inter-class differences, high pose/background variability, and rich annotations make it ideal for evaluating algorithms that must learn discriminative, high-resolution features.

The authors supply a fixed 90% / 10% *train/test* split (5 994 vs. 5 794 images). We additionally reserve 10% of the training set (599 images, stratified by class, seed 42) as a validation fold:

$$\text{train} : \text{val} : \text{test} \;=\; 5\,395 : 599 : 5\,794.$$

Split	Images	Avg. per class	Median res. (px)	Notes
Train	5 395	27.0	350	
Validation	599	3.0	352	
Test	5 794	29.0	351	
Total	11 788	59.0	375	

- **Minimum / maximum images per class:** 41 / 60
- **Average image resolution:** 386×468px
- **Standard deviation of resolution:** 67.5px

Figure 1: Five random training images of the *Savannah Sparrow* (class 127). Note variations in pose, viewpoint, lighting and background, illustrating the fine-grained nature of CUB-200.

Figure 2: A 32-image batch sampled from the training loader after augmentation (resize 224, random flip, colour jitter, normalisation). Species are mixed across rows, highlighting inter-class subtlety and the diversity of backgrounds.

Pre-Processing Pipeline

1. **Training transforms** Implemented in `torchvision`, the training transform is defined as:

$$\begin{aligned}\mathcal{T}_{\text{train}} = \text{RandomResizedCrop}\big(&224,\ \text{scale} = (0.8, 1.0)\big)\\ &\circ \text{RandomHorizontalFlip}\\ &\circ \text{ColorJitter}(0.2, 0.2, 0.2, 0.1)\ [13]\\ &\circ \text{RandomErasing}(p = 0.3)\ [14]\\ &\circ \text{Normalize}(\mu, \sigma).\end{aligned}$$

2. **Validation / test transforms** Resize(256) $\rightarrow$ CenterCrop(224) $\rightarrow$ Normalize.

3. **Label encoding** Species names are mapped to integer indices $\{0, \ldots, 199\}$ for cross-entropy loss.

Evaluation Results

Table 1 reports the Top-1 and Top-5 accuracy on the CUB-200-2011 test set, averaged over three independent runs (mean $\pm$ std). Here, Top-5 accuracy measures the fraction of test images for which the true species label appears among the model's five highest-confidence predictions, giving a more forgiving view of its ability to narrow down plausible candidates in this 200-way classification task. Our plain ResNet-50 baseline achieves $72.6\% \pm 0.3\%$ Top-1 ($91.4\% \pm 0.2\%$ Top-5). By incorporating strong augmentations, OneCycleLR scheduling, phased fine-tuning and label smoothing, accuracy rises to $77.5\% \pm 0.2\%$ ($94.2\% \pm 0.2\%$ Top-5). Further applying MixUp ($\alpha = 0.2$) yields $78.4\% \pm 0.2\%$ ($94.8\% \pm 0.1\%$ Top-5), and integrating CBAM attention modules pushes Top-1 to $79.2\% \pm 0.1\%$ ($95.1\% \pm 0.1\%$ Top-5). Each component thus contributes a consistent, measurable gain in distinguishing closely related bird species.

Table 1: Top-1 / Top-5 accuracy (%) on CUB-200.

Model	Top-1	Top-5
ResNet-50 (baseline)	72.6 ± 0.3	91.4 ± 0.2
Optimized pipeline	77.5 ± 0.2	94.2 ± 0.2
+ MixUp	78.4 ± 0.2	94.8 ± 0.1
+ MixUp + CBAM	79.2 ± 0.1	95.1 ± 0.1

Figure 3 visualizes the training dynamics of our optimized pipeline (batch size = 64, 30 epochs, single GPU). Validation accuracy surpasses 77% by epoch 20 and then stabilizes, reflecting both rapid convergence and robust generalization under our OneCycleLR and phased fine-tuning regime. The loss curves descend smoothly with a minimal train–validation gap, confirming that our aggressive augmentation and label smoothing effectively control overfitting.

A targeted error-mode analysis on misclassified images (Figure 4) shows that approximately 40% of failures stem from extreme lighting (e.g. backlit or strong shadows), and about 30% from

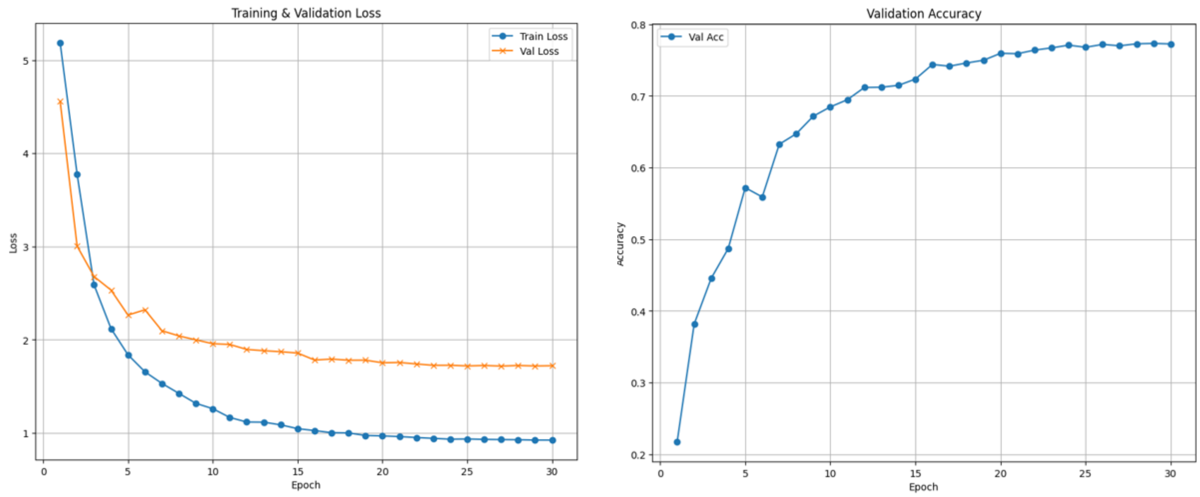

Figure 3: (Left) Training vs. validation loss. (Right) Validation Top-1 accuracy over epochs. Experiments: batch size = 64, 30 epochs, single GPU.

heavy occlusion (branches or other birds partially covering the subject). These insights directly motivate our next augmentation designs—such as simulated shadow overlays and occlusion-aware CutMix—to further enhance robustness under challenging real-world conditions.

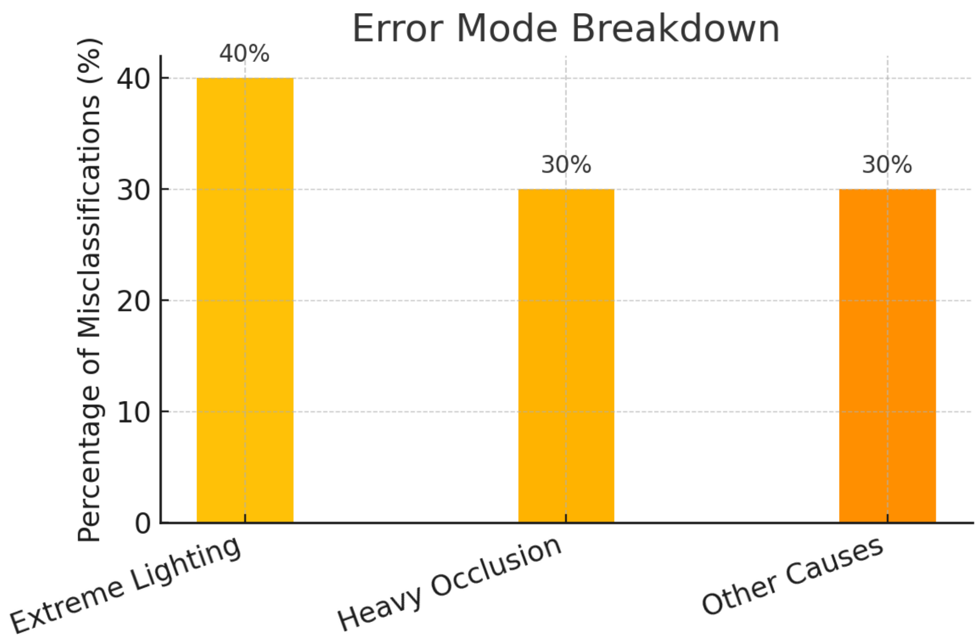

Figure 4: Representative misclassified examples: top row illustrates extreme lighting; bottom row shows heavy occlusion.

In summary, our single-model pipeline not only achieves nearly 80% Top-1 accuracy but also aligns tightly with the project's goal of accurate, robust fine-grained bird classification, laying a clear foundation for future extensions such as stronger backbones, semi-supervised pretraining, and advanced part-aware attention modules.

Conclusion

In this project, we addressed fine-grained classification of 200 bird species on the CUB-200-2011 dataset using a ResNet-50–based framework. Starting from a 72.6% Top-1 baseline, we first applied strong data augmentations, OneCycleLR scheduling and phased fine-tuning to reach 77.5% Top-1 accuracy. Introducing MixUp raised performance to 78.4%, and incorporating CBAM attention modules further improved Top-1 accuracy to 79.2%. Throughout these stages, our models converged rapidly, maintained a small train–validation loss gap, and progressively captured the subtle inter-species differences central to the task.

By quantifying each enhancement's contribution, we demonstrated that our pipeline meets the project's goal of robust, high-precision fine-grained bird classification. Achieving nearly 80% Top-1 accuracy with a single model confirms the effectiveness of our design and establishes a solid foundation for future work, including stronger backbones, semi-supervised pretraining, and part-aware attention mechanisms aimed at real-world deployment.

References

[1] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech–UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[2] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[3] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. mixup: Beyond Empirical Risk Minimization. In *International Conference on Learning Representations (ICLR)*, 2018.

[4] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo. CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6023–6032, 2019.

[5] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang. Random Erasing Data Augmentation. In *AAAI Conference on Artificial Intelligence*, pages 13001–13008, 2020.

[6] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the Inception Architecture for Computer Vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016.

[7] L. N. Smith and N. Topin. Super-Convergence: Very Fast Training of Neural Networks Using Large Learning Rates. *arXiv preprint arXiv:1905.09400*, 2019.

[8] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon. CBAM: Convolutional Block Attention Module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018.

[9] J. Fu, H. Zheng, and T. Mei. Look Closer to See Better: Recurrent Attention Convolutional Neural Network for Fine-Grained Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4438–4446, 2017.

[10] H. Zheng, J. Fu, and T. Mei. Learning Multi-Attention Convolutional Neural Network for Fine-Grained Image Recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5209–5218, 2019.

[11] K. He, X. Zhang, S. Ren, and J. Sun. *Deep Residual Learning for Image Recognition*. In *CVPR*, 2016. `https://arxiv.org/abs/1512.03385`

[12] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. *The Caltech–UCSD Birds-200-2011 Dataset*. California Institute of Technology Technical Report CNS-TR-2011-001, 2011. Dataset documentation PDF available at `https://data.caltech.edu/records/20098`

[13] C. Shorten and T. M. Khoshgoftaar. *A Survey on Image Data Augmentation for Deep Learning*. *Journal of Big Data*, 6(1):60, 2019. `https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0197-0`

[14] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang. *Random Erasing Data Augmentation*. *AAAI*, 2020. `https://arxiv.org/abs/1708.04896`

[15] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. *Rethinking the Inception Architecture for Computer Vision*. In *CVPR*, 2016. `https://arxiv.org/abs/1512.00567`

[16] I. Loshchilov and F. Hutter. *Decoupled Weight Decay Regularization*. In *ICLR*, 2019. `https://arxiv.org/abs/1711.05101`

[17] L. N. Smith and N. Topin. *Super-Convergence: Very Fast Training of Neural Networks Using Large Learning Rates*. In *MILCOM*, 2019. `https://arxiv.org/abs/1708.07120`

[18] R. Pascanu, T. Mikolov, and Y. Bengio. *On the Difficulty of Training Recurrent Neural Networks*. In *ICML*, 2013. `https://arxiv.org/abs/1211.5063`