Enhancing Bird Species Classification with CNNs and Vision Transformers: A Comparative Study using Transfer Learning Approach

Ya-Tzu Lee
Georgia Institute of Technology
ylee646@gatech.edu

Qingyang Liang
Georgia Institute of Technology
qliang46@gatech.edu

Zhehui Liu
Georgia Institute of Technology
zliu715@gatech.edu

Ran Ran
Georgia Institute of Technology
rran6@gatech.edu

Abstract

Bird classification plays a crucial role in various ecological and environmental studies. In this paper, we present a comprehensive investigation into developing a robust and efficient bird classification model, employing Convolutional Neural Network (CNN) architectures and the Vision Transformer (ViT) model. By incorporating methodologies such as transfer learning, data augmentation, and model fine-tuning, we construct a model with high accuracy and resilience to perturbations. Furthermore, we extend our exploration to model compression through distilled learning and model generalization via the creation of a few-shot learning scenario. The outcomes illuminate that both enhanced CNN architecture and the ViT model yield comparable performance, with different underlying architectures. Through integration of the cutting-edge techniques and utilization of CNNs and the ViT model, our study advances the fine-grained bird classification methodologies and carries practical implications for real-world bird classification applications.

1. Introduction

Recently, the field of computer vision has gained remarkable progress in image classification tasks due to the development of deep learning techniques. The fine-grained bird classification problem has been catching the eyes of researchers as a challenge since there are vast numbers of bird species and the subtle visual differences between the species are hard to deal with. In this study, we explore the fine-grained bird classification problem with a focus on studying the effectiveness and robustness of various Convolutional Neural Network (CNN) models and the Vision Transformer (ViT) model in different test environments. We also aim at developing a model which is capable of accurately classifying the bird species with a limit number of labeled samples.

To achieve our objective, we train four popular CNN models, namely, ResNet [7], VGG [21] , MobileNet [8], and EfficientNet [24], on the original base dataset and evaluate their performances on the augmented test datasets. The four CNN models are proved to have good results in bird classification field [20, 10, 6, 11], with ResNet renowned for its ability to effectively train very deep networks by utilizing skip connections or residual blocks, VGG known for its simplicity by utilizing stack of small size convolutional filters with fixed structure for feature extraction, MobileNet noted for its lightweight and low computational complexity designed for mobile vision applications, and EfficientNet known for its balanced architecture by efficiently scaling the depth, width, and resolution together to achieve superior performance. Then we fine-tune the best architecture by modifying the hyper-parameters, loss function, learning policy, and image segmentation to further enhance its performance. Moreover, we explore the potential of the state-of-the-art ViT model and compare it against the CNN models. In addition, a challenging few-shot learning scenario is created by introducing new bird species for model evaluation, and distilled learning is implemented for exploring model compression to fit the model into real world tasks.

Currently, the development of artificial intelligence promotes the bird species classification problem in acoustic [12] and visual features of birds, while in this study we focus on the visual features recognition side. Based on the developing CNN architectures, researchers explore diverse methodologies for bird species recognition. Kondaveeti et al. [13] introduces a transfer learning approach, which is a common practice in the absence of large labeled datasets, utilizing MobileNetV2 for bird species recognition. The study reveals that the model's generalization is hindered by intricate visual differences among closely related bird

species. Liu et al. [15] introduces TransIFC, a method that incorporates invariant cues-aware feature concentration learning for efficient fine-grained bird classification. The study shows the importance of capturing invariant features for handling variations in the same bird species. Wang et al. [27] introduces a novel approach by leveraging attention mechanisms and decoupled knowledge distillation. The model is enabled to focus on relevant details and effectively learn knowledge from a teacher model.

Despite all the innovative approaches, current practice in fine-grained bird classification still faces limitations. Limited labeled data, especially for less common bird species, poses a significant obstacle to the development of robust models. In addition, capturing fine-grained visual cues which discriminate closely related bird species in various background noises remains a challenging task for traditional deep learning techniques.

The bird species classification problem has various practical applications, including ornithology, wildlife conservation, ecological research, and so on. Improving the accuracy in bird species classification can help researchers track and monitor bird populations, better understand and maintain ecological balance. Our research can provide valuable insights into the strengths and weaknesses of modern computer vision models for researchers to seek the most suitable model in their specific bird classification tasks. Moreover, by exploring the model with few-shot learning and distilled learning techniques, our research shows the model's capacity to be generalized effectively with limited training data, which is important for real-world applications where labeled data for certain bird species may be insufficient.

For the purpose of this study, we randomly sample 200 bird species from the BIRDS 525 SPECIES dataset to create the base training set due to computational constraints. In addition, we use 50 new species upon the base training dataset in the few-shot learning evaluation setup. To ensure model generalization and robustness, we augmented the training and testing datasets at different stages with data augmentation techniques, and the details are described in section 2.1.2.

2. Approach

In our approach, we set out to enhance bird species classification by leveraging transfer learning with pre-trained CNN models and exploring the potential of vision transformers. Our primary objectives encompassed achieving accurate bird species classification, even in few-shot learning scenarios, while ensuring the robustness of our approach in the face of real-life data augmentation challenges, such as noise and blurriness. Additionally, we aim to maintain an efficient training time for the models. To accomplish these goals, we conducted a series of experiments and optimizations, building upon existing pre-trained models while introducing innovative optimization and enhancement procedures. The entire process can be summarized as shown in Figure 1.

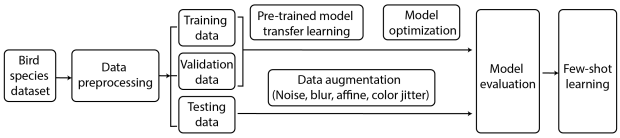

Figure 1. Overview of the experimental approach.

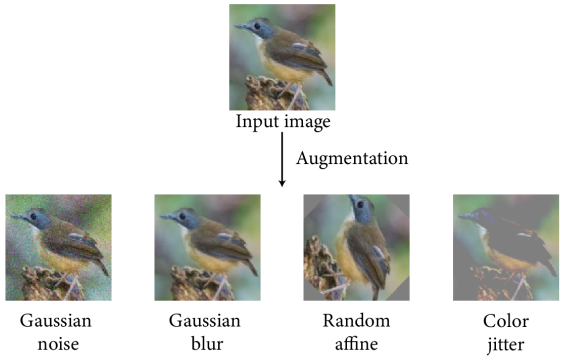

Figure 2. Examples of data augmentations.

2.1. Workflow

2.1.1 Data collection

We acquired the bird classification dataset from Kaggle [https://www.kaggle.com/datasets/gpiosenka/100-bird-species], containing 224x224x3 color images of 525 bird species. The dataset was split into 84,635 training images, 2,625 validation images and 2,625 test images. This dataset is with high quality of images, similar to the classic bird dataset CUB-200 [26], making it suitable for fine-grained image classification. We selected 200 out of 525 total classes to reduce the computational cost for training and leave room for experiments with new classes.

2.1.2 Data preprocessing

The images used for the experiments were all preprocessed by resizing and normalizing. To simulate real-life scenarios in bird image acquisition, we carefully selected several data augmentations, including Gaussian blur, random affine transformations, and color jitter, along with custom noise augmentations like Gaussian noise, salt and pepper noise, and speckle noise, as examples shown in Figure 2. These augmentations were carefully chosen to mimic real-world scenarios where bird images may be affected by blurriness

caused by motion, varying orientations and scales at different focal lengths, different exposure levels, and noises at low light conditions.

2.1.3 Models training and evaluation

We began by training four pre-trained models obtained from Torchvision, namely ResNet18, VGG19, MobileNet-v3-small, and EfficientNet-b0, using transfer learning. The training process involved 50 epochs, and we evaluated the models based on cross-entropy losses and training accuracy. EfficientNet-b0 demonstrated superior performance compared to the other models, aligning with its known efficiency, performance, transfer learning capability and robustness to data augmentation [25]. Notably, all models converged well before the 50-epoch limit, prompting us to use fewer epochs in subsequent training stages to enhance computational efficiency. We also experimented with training the models on datasets with combinations of all devised augmentations, observing improved testing accuracy under various augmentation scenarios when compared to training on the base dataset, at a cost of a much slower training speed as expected. However, training on all possible augmented scenarios may be impractical, leading us to focus on optimizing the base model trained on the original dataset.

2.1.4 Model fine-tuning and investigation

To further enhance the EfficientNet-b0 model, we explored hyperparameter tuning, including changing optimizers, learning rates, and learning steps. Additionally, we implemented advanced techniques such as Label Smoothing [18] and OneCycle learning policy [22] to optimize the loss function and learning rate, leading to an improved EfficientNet model for further experiments. We also investigated the use of image segmentation for feature extraction and distilled learning with teacher-student models to expedite the learning process. Inspired by applications of the Vision Transformer (ViT) model to fine-grained classification tasks [1], we experimented with ViT for bird species classification, fine-tuning the model by updating only the final layer parameters or allowing all layers to be adjusted. The ViT model demonstrated comparative results with the best-performing EfficientNet-b0 model, while exhibiting enhanced robustness to all data augmentations.

2.1.5 Few-shot learning

To simulate real-life scenarios where bird scientists may encounter new bird species with limited labeled images, we conducted few-shot learning experiments. We randomly selected 50 classes that were not present in the previously trained 200 classes, with each class having only 10 images. Our improved EfficientNet model showed promising results in few-shot learning, achieving fast training times (less than a minute) and exhibiting potential for further improvement with state-of-the-art few-shot learning strategies.

2.2. Code implementation

We based our work on the ResNet training notebook using PyTorch, and made significant modifications to tailor it to our specific needs. Our modifications included the addition of 1) specified class selection 2) custom data augmentations 3) configuration variables for hyper-parameter fine-tuning 4) multiple pre-trained model implementations 5) image segmentation 6) distilled learning 7) few-shot learning simulation.

3. Experiments and Results

3.1. CNN models and EfficientNet

In this section, we first construct a baseline using the traditional and state-of-art CNN models. Comparative studies have been conducted with transfer learning techniques on the following CNN models: EfficientNet, MobileNet, VGG, and ResNet. To compare their performance, we applied the same training method to each CNN model with the following steps:

- **Step 1:** Load the pre-trained CNN models
- **Step 2:** Train the model with base dataset using the same training algorithm (CrossEntropy loss function and step learning rate scheduler have been selected.)

Table 1 shows the comparison of the testing, training, and validation accuracy for four selected CNN models. The EfficientNet has the highest training, validation, and test accuracy among other CNN models. In addition, the augmentation has been conducted to evaluate the CNN model's performance on the images with Gaussian blur, Gaussian noise, and random affine. The results have been demonstrated in Figure 3, which accordingly illustrates that EfficientNet has outstanding performance among the four CNN models. However, as shown in Table 1, EfficientNet requires more training time and computational cost compared with ResNet and MobileNet.

Considering the training outcome, we select EfficientNet as the baseline model. Our main goal is to fine-tune the parameters and hyper-parameters to improve its accuracy and, if possible, reduce its computational cost.

Learning rate is one of the key hyper-parameters during deep learning model training. If learning rate is too large, the model could converge too quickly to a sub-optimal point. If the learning rate is too small, it could overfit the dataset [22]. Two improvements have been selected to avoid overfitting, and enhance overall performance, namely,

CNN Model	Test Acc (%)	Train Acc (%)	Valid Acc (%)	Train Time (s)
VGG19	95.3	99.2	94.7	7319.68
ResNet18	93.3	91.76	93.4	2860.28
MobileNet-v3-small	83.5	74.94	83.4	2882.53
EfficientNet-b0	99.4	97.23	97.8	4347.76

Table 1. The training, validation, test accuracy, and training time for 50 epochs of the CNN models.

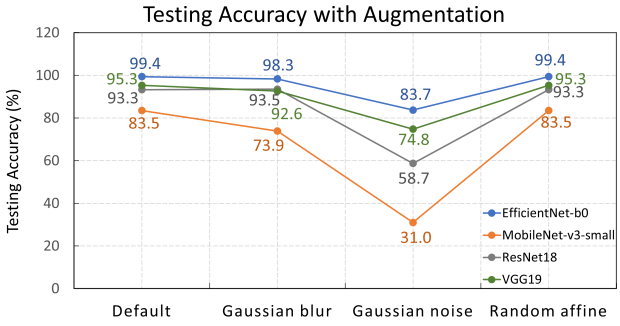

Figure 3. CNN models test accuracy with different augmentation scenarios: 1) default, 2) Gaussian blur, 3) Gaussian noise, 4) random affine.

OneCycle learning policy and Label Smoothing cross entropy.

OneCycle policy changes the learning rate starting from the minimum to the maximum learning rate, then reduce the learning rate to the minimum. The idea is the learning rate increases in the middle of the learning process and acts as a regularization method to keep the model from overfitting [22]. Another method to address overfitting and overconfidence is Label Smoothing [18]. As shown in the Table 2, by using OneCycle learning policy and Label Smoothing cross entropy, the training and validation accuracy of EfficientNet has been improved by approximately 1%. This improvement is impressive as the base accuracy is already at 97%, bringing the model's performance in line with the current state-of-the-art (SOTA) results.

As shown in Figure 3, among all the chosen CNN models, their worst performances are observed during the Gaussian noise scenario. Our hypothesis is that the noise came from the contextual information. Therefore, we propose to improve CNN model accuracy by combining general model training process with image segmentation. One of the benefits of image segmentation is the ability to remove background noise [5], this method is typically applied in self-driving vehicle applications for object detection.

The new workflow of model training has been shown in Figure 4. Instead of training and testing the model with the original dataset, a new block is added for applying a segmentation mask on the original dataset. In this case, the model has been trained by segmented data first. Then, during testing, the test image will be masked by feeding into the segmentation algorithm. The selected segmentation algorithm is FCN ResNet 50 [16], which was pre-trained by COCO [14].

Model	Test Acc (%)	Train Acc (%)	Valid Acc (%)
EfficientNet-b0 (Original)	99.4	97.23	97.8
EfficientNet-b0 (Improved)	99.8	98.90	98.6

Table 2. Comparison between the original and improved EfficientNets: 1) original model with step learning rate and cross entropy loss function; 2) improved model with OneCycle learning policy and Label Smoothing loss function.

The results of the experiment have been shown in Table 3. Surprisingly, the accuracy is slightly worse by masking out the environment. We consider two possible reasons:

- Context information is critical for the bird species classification. It is one of the key features that neural networks learned. Considering correlation between different bird species and the environment, we deem the accuracy reduction reasonable.
- The pre-trained model used during transfer learning could be trained by the image with context already. Feeding in masked images could result in some feature lost and thus, mislead the prediction.

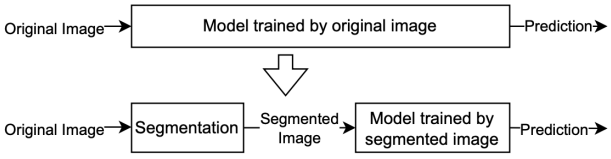

Figure 4. Comparison of model training processes before and after modification.

Figure 5. An example of the original dataset v.s. the segmented dataset.

Model	Test Acc (%)	Train Acc (%)	Valid Acc (%)
Workflow 1	99.8	98.90	98.6
Workflow 2	97.6	97.9	97.0

Table 3. Train, validation and test accuracy comparison of: 1) improved EfficientNet with original workflow (workflow-1); 2) improved EfficientNet with segmentation (workflow-2)

In summary, the EfficientNet has outstanding performance among other traditional CNN models, and its testing accuracy reached 99.8% after our improvements. We highlight the importance of both contextual information and birds' features utilized during the image processing. This model can be used as the benchmark model when exploring other methodologies on bird classification other than CNN.

3.2. Vision Transformers

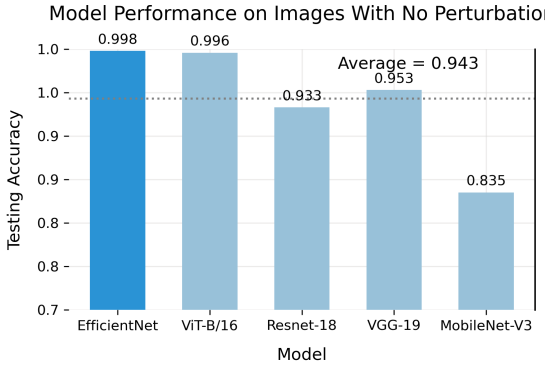

Figure 6. Model performance on default images in test set.

Traditional CNNs struggle with perturbations introduced in test dataset. The best performing CNN model in our experiment, improved EfficientNet, achieves 0.998 accuracy rate with original test images (Figure 6) but only 0.889 accuracy rate with Guassian noises (Figure 7). In addition to CNNs, we also experiment with Transformer-based architectures, in which Vision Transformers demonstrate significantly higher robustness in predicting the test set with noises.

Although none of the noises is introduced during the training process, Vision Transformers can accurately label the test set with unseen noises, achieving a high accuracy rate. Specifically, the vision transformer achieves an unwavering accuracy rate above 0.99 for Gaussian blur, random affine, and speckle noise (Figure 7). This level of performance far surpasses that of traditional CNNs, where the best-performing model in the experiment, EfficientNet, struggles to achieve only 0.889 accuracy rate when dealing with images containing Gaussian noise. The performance of vision transformers in the presence of Gaussian noise, achieving an accuracy rate of 0.989, indicating the ability of Vision Transformers to grasp the underlying patterns and generalizations more effectively than conventional CNN architectures. In terms of model training time for Vision Transformer, it took 1,667.24 seconds and 10 epochs for the training loss to converge, a reasonable training effort.

One key difference between Vision Transformers and CNNs is how they capture global and local information. CNNs use convolutional filters to extract local features, which are then aggregated hierarchically to capture increasingly complex patterns, leveraging the powerful inductive bias of spatial equivariance encoded by convolutional layers. In contrast, Vision Transformers operate almost identically to Transformers used in language, using self-attention, rather than convolution, to aggregate information across locations [2].

Prior work by Raghu et al. [19] examining the feature learning process between CNNs and ViTs demonstrates that the self-attention mechanism allows Vision Transformers to incorporate more global information than traditional CNN architectures, leading to quantitatively different features. Even in the lowest layers of Vision Transformers, self-attention layers have a mix of local heads (small distances) and global heads (large distances), indicating the ability to focus on both local and global patterns. In contrast, CNNs are hard coded to attend only locally in the lower layers .

The ability of Vision Transformers to handle global information enhances their resilience in noisy environments. The self-attention mechanism allows them to effectively filter and focus on relevant global features, mitigating the impact of noise on the final predictions. Consequently, Vision Transformers exhibit an exceptional ability to generalize beyond their training data, making them a promising candidate for robust computer vision applications in real-world scenarios.

3.3. Evaluating model robustness with data augmentation

In this section, we explore the impact of data augmentation on the performance of our model. To investigate the robustness of the model to various augmentations, we applied them to the testing dataset and computed the accuracy score. We hypothesized that certain perturbations, such as adding a bit of noise, might lead to significantly worse performance causing the network to misclassify. Indeed, our result show that this is the case for certain models like ResNet-18 or MobileNet-V3 (Figure 7). However, to our surprise, both EfficientNet-b0 and ViT-B/16 demonstrated remarkable performance even with noise, possibly due to their ability to handle diverse data effectively, making them suitable models for bird classification tasks.

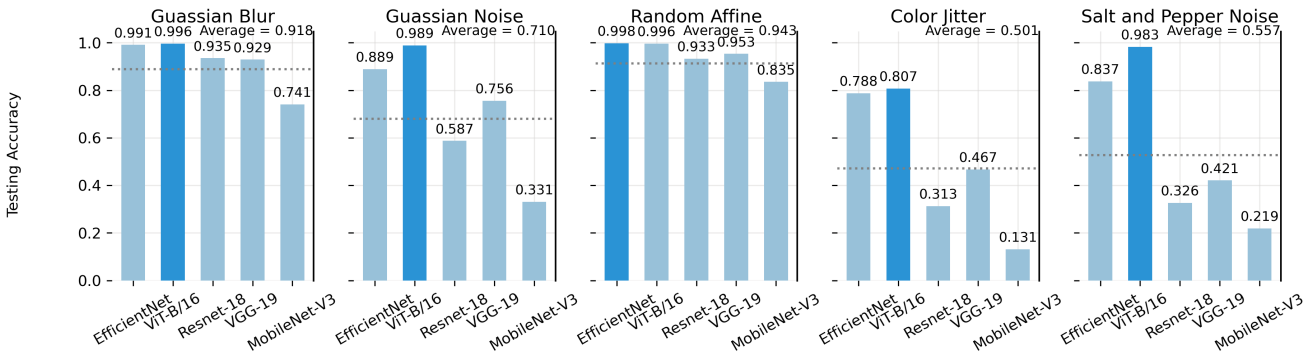

Figure 7. Model performance on images with perturbations.

Moreover, to assess the impact of data augmentation during training, we applied the same augmentations to the training dataset and retrained our best CNN model, EfficientNet-b0. The results demonstrated that the accuracy was nearly 1 for each augmentation when tested on the dataset. This indicated that the model successfully learned to generalize from augmented data during training, highlighting the usefulness of data augmentation, especially when the training data size is limited. Therefore, exploring data augmentation techniques in semi-supervised learning, like FixMatch [23], or considering image generation methods such as GANs [4], holds promise in enhancing training outcomes when dealing with limited datasets.

3.4. Distilled learning

Noticing the trade-off between model training time and model performance, we conduct distilled learning to explore model compresssion. In our approach to distilled learning, we utilize the enhanced EfficientNet architecture as the teacher model and MobileNet as the student model which maintains half the size of the teacher model. The process involves a combination of soft and hard loss components, resulting in an enhanced student model. This method demonstrates an improvement in accuracy to 0.925, 0.835, and 0.925 under default, Gaussian blur, and random affine augmentations, respectively. This compelling result points the direction to follow similar principles for the exploration of lightweight models suitable for real-world applications, such as ShuffleNet [17] and SqueezeNet [9].

3.5. Few-Shot learning

In real-world scenarios, where obtaining sufficient training data is challenging, as often happens when discovering new bird species, few-shot learning becomes a crucial aspect. In this context, we investigated few-shot learning for our model, using 50 new classes, each containing only 10 training images. Due to time constraints, we focused on the base model for this analysis. Surprisingly, when tested on our base model, the testing accuracy achieved an impressive score of 0.916, despite the model suffering from overfitting (with a training accuracy of 0.998 and a validation accuracy of 0.912).

We also attempted to mitigate overfitting using regularization methods, such as dropouts, and improve the classifier by replacing the final linear layer classifier with cosine similarity scoring. However, these attempts showed little improvement, with some decline in performance, as calculating cosine similarity may not necessarily adapt the model to recognize entirely new classes. Further experimentation, involving more advanced techniques, such as Model-Agnostic Meta-Learning (MAML) by learning the gradient descent using LSTM [3], may be needed to address the challenges of few-shot learning effectively.

4. Conclusion

In summary, we conducted a thorough examination of state-of-the-art CNN models for bird classification and determined that the EfficientNet model, combined with OneCycle Learning Policy and Label Smoothing, performed best. We also delved into the potential of image segmentation and distilled learning to enhance model performance. Additionally, we discovered that a Vision Transformer model also demonstrated comparable performance and was more robust to perturbations, making it an attractive alternative. These models exhibited robustness to real-life data augmentations, making them highly suitable for practical applications. These models can be compressed via distilled learning technique, and can be effectively employed for real-life few-shot learning scenarios, though further improvements on few-shot learning or exploration on generative models for increasing the dataset size could be carried out. With our comprehensive comparison and investigation of deep learning models for bird classification, we believe our model can serve as a valuable resource for the bird scientist community.

References

[1] Marcos V. Conde and Kerem Turgutlu. Exploring vision transformers for fine-grained classification, 2021.

[2] A. Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[3] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks, 2017.

[4] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[5] Yanming Guo, Yu Liu, Theodoros Georgiou, and Michael Lew. A review of semantic segmentation using deep neural networks. *International Journal of Multimedia Information Retrieval*, 7, 06 2018.

[6] Yulius Harjoseputro, Ign Yuda, Kefin Pudi Danukusumo, et al. Mobilenets: Efficient convolutional neural network for identification of protected birds. *IJASEIT (International Journal on Advanced Science, Engineering and Information Technology)*, 10(6):2290–2296, 2020.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[8] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019.

[9] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.

[10] Shazzadul Islam, Sabit Ibn Ali Khan, Md Minhazul Abedin, Khan Mohammad Habibullah, and Amit Kumar Das. Bird species classification from an image using vgg-16 network. In *Proceedings of the 7th International Conference on Computer and Communications Management*, pages 38–42, 2019.

[11] Aldi Jakaria and Hilman Ferdinandus Pardede. Comparison of classification of birds using lightweight deep convolutional neural networks. *Jurnal Elektronika dan Telekomunikasi*, 22(2):87–94, 2022.

[12] Stefan Kahl, Mary Clapp, W Alexander Hopping, Hervé Goëau, Hervé Glotin, Robert Planqué, Willem-Pier Vellinga, and Alexis Joly. Overview of birdclef 2020: Bird sound recognition in complex acoustic environments. In *CLEF 2020-Conference and Labs of the Evaluation Forum*, volume 2696, 2020.

[13] Hari Kishan Kondaveeti, SVN Sai Vignesh Guturu, KS Jayan Praveen, and Samparthi VS Kumar. A transfer learning approach to bird species recognition using mobilenetv2. In *2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 787–794. IEEE, 2023.

[14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[15] Hai Liu, Cheng Zhang, Yongjian Deng, Bochen Xie, Tingting Liu, Zhaoli Zhang, and You-Fu Li. Transifc: invariant cues-aware feature concentration learning for efficient fine-grained bird image classification. *IEEE Transactions on Multimedia*, 2023.

[16] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation, 2015.

[17] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, pages 116–131, 2018.

[18] Rafael Müller, Simon Kornblith, and Geoffrey Hinton. When does label smoothing help?, 2020.

[19] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy. Do vision transformers see like convolutional neural networks?, 2022.

[20] Kazi Md Ragib, Raisa Taraman Shithi, Shihab Ali Haq, Md Hasan, Kazi Mohammed Sakib, and Tanjila Farah. Pakhichini: Automatic bird species identification using deep learning. In *2020 Fourth world conference on smart trends in systems, security and sustainability (WorldS4)*, pages 1–6. IEEE, 2020.

[21] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[22] Leslie N. Smith. A disciplined approach to neural network hyper-parameters: Part 1 – learning rate, batch size, momentum, and weight decay, 2018.

[23] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence, 2020.

[24] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, pages 10096–10106. PMLR, 2021.

[25] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks, 2020.

[26] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. Caltech-ucsd birds-200-2011 (cub-200-2011). Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[27] Kang Wang, Feng Yang, Zhibo Chen, Yixin Chen, and Ying Zhang. A fine-grained bird classification method based on attention and decoupled knowledge distillation. *Animals*, 13(2):264, 2023.