Proceedings of CONF-MLA 2025 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2025.LD27853

Comparative Study of Zero-shot and Fine-tuned Vision Models--Evaluating CLIP, LiT, and Swin Transformer on Fine-grained Bird Classification

YihuaWang

Wuhan Polytechnic University, Wuhan, China
2786250363@qq.com

Abstract. Fine-grained image classification is challenging because categories are often separated by only minor visual cues, requiring models to capture very fine details for accurate discrimination. The latest advances in vision-language models, such as CLIP and LiT, have demonstrated strong zero-shot performance on general image recognition tasks, but their effectiveness in fine-grained domains remains underexplored. In this study, we conduct a comparative evaluation of CLIP, LiT, and the vision-only Swin Transformer on the CUB-200-2011 bird dataset. For zero-shot classification, we assess CLIP and LiT using a consistent prompt template, while for fine-tuning, we train both CLIP and Swin end-to-end using the AdamW optimizer. Results show that LiT outperforms CLIP in zero-shot settings (63.96% vs. 51.55% Top-1 accuracy), while Swin achieves the highest performance after fine-tuning (83.47% Top-1 accuracy). These findings highlight a trade-off between generalization and fine-grained specialization, and suggest that future work should explore lightweight adaptation techniques to bridge the performance gap without sacrificing zero-shot flexibility.

Keywords: Zero-shot learning, Fine-tuning, Vision-language models, CLIP, Swin Transformer.

1. Introduction

Recently, zero-shot learning, which enables classification without having seen the visual classes during training, has emerged as a new paradigm in vision-language models [1]. Two widely studied zero-shot models, CLIP [1] and LiT (Locked-image Tuning) [2], have demonstrated strong performance across a wide range of open-domain visual tasks by aligning images and textual descriptions in a shared embedding space. While these models perform well on general image recognition tasks, it is unclear how well they handle fine-grained categories with subtle differences. Such tasks typically require high-resolution visual discrimination, which is traditionally tackled through fully supervised fine-tuning of high-capacity vision models, such as Swin Transformer [3].

This gap in understanding motivates us to investigate whether zero-shot models like CLIP and LiT can match the performance of fully fine-tuned models when applied to fine-grained classification tasks. In particular, we aim to understand how different architectures—especially

DOI: 10.54254/2755-2721/2025.LD27853

multimodal models such as CLIP and LiT, compared to vision-only architectures like Swin Transformer—transfer to specialized domains that require detailed visual discrimination.

To explore this, we conduct a comprehensive comparative study using the CUB-200-2011 dataset [4], a commonly adopted benchmark for fine-grained species classification. Our evaluation covers both the zero-shot performance of CLIP and LiT, and the supervised fine-tuning performance of CLIP and Swin Transformer. Through this comparison, we aim to uncover the trade-offs between generalization and specialization, and to offer practical insights into how different model types can be adapted for real-world fine-grained recognition tasks.

2. Related work

Recent years vision-language models have made a significant progress, most notably with the introduction of CLIP by openAI [1], which aligns image and text embeddings through contrastive learning on large-scale web data. CLIP have demonstrated remarkable generalization abilities, particularly in zero-shot classification, where it can classify images it has never seen during training by leveraging natural language descriptions.

LiT , proposed by Google [2], further improves zero-shot transfer performance by freezing the image encoder and only tuning the text tower. This design choice reduces overfitting and enhances the stability of representation alignment, leading to stronger performance on unseen categories with minimal computation overhead

In contrast to these multimodal models, Swin Transformer [3] represents a purely visual, hierarchical transformer architecture that delivers strong performance across several standard vision tasks. However, Swin does not support zero-shot inference by default and requires full supervised fine-tuning to adapt to specific downstream tasks.

The dataset used in our study, CUB-200-2011 [4], is a fine-grained bird classification benchmark that includes 200 species and over 11,000 images with detailed annotations. It poses a unique challenge due to the subtle visual differences between classes, making it ideal for evaluating both generalization and specialization capabilities of different models.

Finally, our study is grounded in the broader context of transfer learning paradigms—particularly the comparison between zero-shot transfer, which relies on pretrained alignment with language, and fine-tuning, which adapts models directly to the target dataset using labeled supervision. Each approach presents different trade-offs in terms of scalability, flexibility, and accuracy in fine-grained classification tasks.

3. Method

3.1. Dataset

We use the CUB-200-2011 dataset [4]. Each image is labeled with its corresponding category. We follow the official data split, dividing the dataset evenly into training and testing sets (50% each). Only the image-level class labels are used for experiments.

3.2. Zero-shot evaluation (CLIP & LiT)

For CLIP [1] and LiT [2], we perform zero-shot classification by embedding both images and class names into a shared embedding space and computing cosine similarity.

Text Prompts: To ensure consistency, we use the same textual template for both CLIP and LiT: “a photo of a {class name} bird”. For each class in the CUB-200-2011 dataset, we tokenize the prompt

DOI: 10.54254/2755-2721/2025.LD27853

and encode it using the model's text encoder to generate class-level text embeddings for zero-shot prediction.

Image Preprocessing: Images are resized and center-cropped to 224×224 pixels, then normalized according to each model's requirements.

Feature Extraction: We extract image and text features using pretrained encoders without fine-tuning.

Prediction: Top-1 and Top-5 accuracy are computed by ranking the similarity scores and selecting the highest-scoring classes.

Evaluation: Accuracy and confusion matrices are computed based on the predicted Top-1 class.

3.3. Fine-tuning setup

For both CLIP [1] and Swin Transformer [3], we adopt a full-model fine-tuning strategy, updating all trainable parameters. Optimization is performed using the AdamW optimizer with a learning rate of 1×10^{-4} and weight decay of 1×10^{-2}, ensuring effective convergence while mitigating overfitting. Training spans 5 epochs, with batch sizes of 32 for Swin and 64 for CLIP. To prevent overfitting and strengthen generalization, we adopt data augmentation operations such as random horizontal flips and random rescaling. For CLIP, the image encoder, text encoder, and projection layers are all unfrozen and optimized jointly, while Swin Transformer is trained end-to-end from ImageNet-pretrained weights. Cross-entropy loss is minimized with respect to ground-truth labels, and model selection is based on Top-1 accuracy on the test split after the final epoch.

3.4. Evaluation metrics

We report:

Top-1 Accuracy: Percentage of correctly classified images

Top-5 Accuracy: Proportion of samples where the correct label is among the top 5 predictions

Confusion Matrix: Visualized on the top 10 most frequent classes to analyze class-wise errors

4. Results and discussion

Our experiments on the CUB-200-2011 dataset reveal distinct strengths and limitations across zero-shot and fine-tuning paradigms. In the zero-shot setting, LiT achieved a Top-1 accuracy of 63.96%, outperforming CLIP's 51.55%, indicating that LiT's locked-image tuning better preserves semantic alignment for fine-grained categories. However, both models struggled to differentiate visually similar bird species, as shown in their confusion matrices, with CLIP exhibiting more frequent misclassifications among morphologically close classes such as albatrosses.

Fine-tuning substantially improved performance for both architectures. CLIP's accuracy increased to 62.86%, while Swin Transformer reached 83.47% Top-1 accuracy, the highest among all models. This performance gap suggests that vision-only transformers, when trained end-to-end, can better capture subtle visual details critical for fine-grained recognition, whereas vision-language models may require specialized adaptation strategies to match that level of precision.

As shown in Table 1, LiT achieves the highest zero-shot Top-1 accuracy (63.96%), while Swin Transformer achieves the best performance after fine-tuning (83.47% Top-1 accuracy). The results indicate a clear advantage of fine-tuning for vision-only models in fine-grained tasks.

DOI: 10.54254/2755-2721/2025.LD27853

Table 1. Performance comparison of zero-shot and fine-tuned models

Model	Type	Top-1 ACC	Top-5 ACC
CLIP	Zero-shot	51.55%	81.95%
LiT	Zero-shot	63.96%	90.77%
CLIP	Fine-tune	62.86%	90.14%
Swin	Fine-tune	83.47%	96.27%

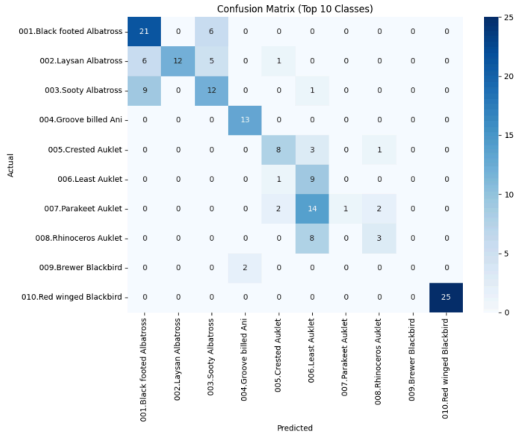

Figure 1. CLIP zero-shot

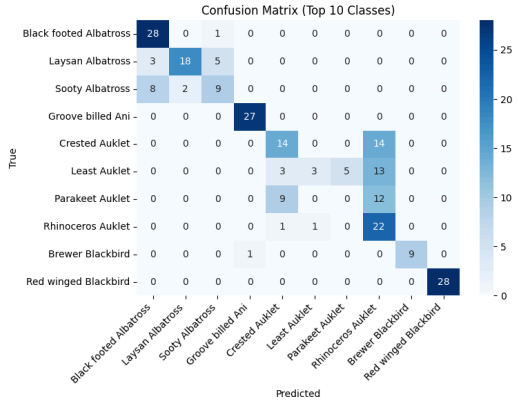

Figure 2. LiT zero-shot

DOI: 10.54254/2755-2721/2025.LD27853

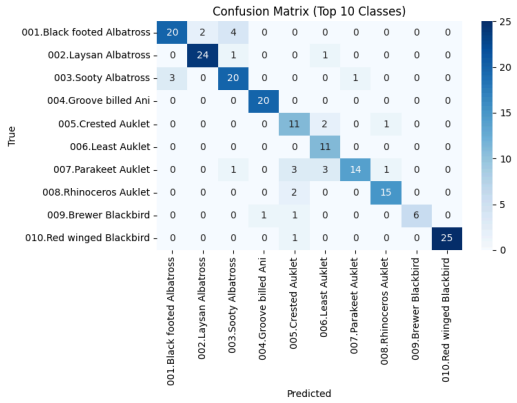

Figure 3. CLIP fine-tune

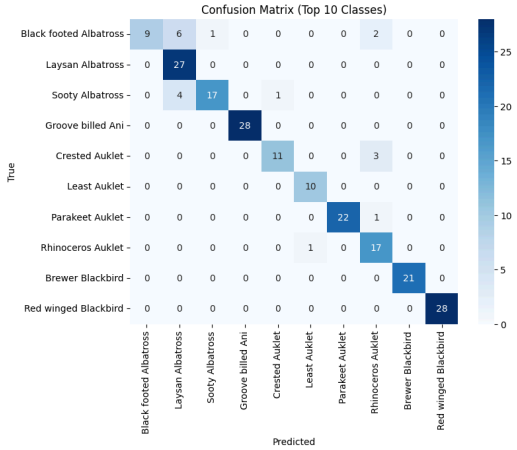

Figure 4. Swin fine-tune

Figure 1 shows the confusion matrix for CLIP in the zero-shot setting, revealing frequent misclassifications among visually similar bird species. As shown in Figure 2, LiT's zero-shot confusion matrix indicates better discrimination capability compared to CLIP. After fine-tuning, the confusion matrix of CLIP (see Figure 3) shows improved but still limited class separation. In contrast, Figure 4 demonstrates that Swin Transformer after fine-tuning achieves much clearer diagonal dominance, indicating significantly better fine-grained recognition performance.

5. Conclusion

This study demonstrates that while zero-shot vision-language models like CLIP [1] and LiT [2] are valuable for classification without labeled data, their performance is limited in fine-grained domains such as bird species recognition. Fine-tuning remains a powerful method for improving accuracy, with the Swin Transformer [3] achieving the best results. However, this comes with increased computational and data requirements. Future research should focus on bridging this gap by developing lightweight adaptation techniques—such as prompt tuning [6-8] or adapter-based methods [5] —that enhance fine-grained recognition capabilities without sacrificing the data efficiency and generalization benefits of zero-shot learning.

DOI: 10.54254/2755-2721/2025.LD27853

Acknowledgments

First and foremost, I would like to express my profound gratitude to my supervisor, Prof. David Woodruff. Throughout this research, he has provided me with invaluable guidance, constructive feedback, and unwavering support. His insights and encouragement were instrumental and essential in shaping this work.

References

[1] A. Radford et al., “Learning Transferable Visual Models From Natural Language Supervision, ” in Proc. ICML, 2021, pp. 8748–8763.
[2] X. Zhai et al., “LiT: Zero-Shot Transfer with Locked-image Text Tuning, ” in Proc. CVPR, 2022, pp. 18123–18133.
[3] Z. Liu et al., “Swin Transformer: Hierarchical Vision Transformer using Shifted Windows, ” in Proc. ICCV, 2021, pp. 10012–10022.
[4] C. Wah et al., “The Caltech-UCSD Birds-200-2011 Dataset, ” Caltech, Tech. Rep. CNS-TR-2011-001, 2011.
[5] A. Frome et al., “DeViSE: A Deep Visual-Semantic Embedding Model, ” in Proc. NeurIPS, 2013, pp. 2121–2129.
[6] X. Liu et al., “Patch-Prompt Aligned Bayesian Prompt Tuning for Vision-Language Models, ” in Proc. UAI, 2024, pp. 2309–2330.
[7] S. Jie et al., “Memory-Space Visual Prompting for Efficient Vision-Language Fine-Tuning, ” in Proc. ICML, 2024, vol. 235, pp. 22062–22074.
[8] L. Lan et al., “Efficient Prompt Tuning of Large Vision-Language Model for Fine-Grained Ship Classification, ” arXiv preprint arXiv: 2403.08271, 2024.