Fine-Grained Mushroom Classification with Prompt-Enhanced Vision–Language Models

Valeria Cerciello
University of Zurich
valeria.cerciello@uzh.ch

Harshita Gupta
University of Zurich
harshita.gupta@uzh.ch

Diana Korotun
University of Zurich
diana.korotun@uzh.ch

Shuai Wang
University of Zurich
shuai.wang@uzh.ch

Abstract

Fine-grained visual classification is challenging in ecological studies because many species look very similar, and the number of labeled images is often small. In this work, we examine how well vision-language models, specifically CLIP, can classify mushroom species. We construct a balanced dataset of 169 mushroom species and evaluate several OpenCLIP backbones using different text prompt designs and limited amounts of visual supervision. To measure performance, we compare CLIP with supervised ResNet models in three settings: (i) a baseline trained on all species and all images, (ii) a model trained on only 50% of the species (classes), and (iii) a model trained on only 50% of the images for each species. We further explore how to adapt CLIP with limited labeled data using Low-Rank Adaptation (LoRA), which enables efficient fine-tuning by updating only a small number of parameters.

Our results show that CLIP consistently outperforms the supervised ResNet baselines across all evaluation settings. On the full dataset, our best CLIP configuration reaches 95.5% Macro-F1 compared to 80.8% for a fully supervised ResNet (about $1.2\times$ higher). Under strong data constraints, the gap widens dramatically: when the ResNet is trained on only 50% of the species or 50% of the images, its Macro-F1 drops below 0.02%, whereas a zero-shot CLIP model still achieves about 7.1% Macro-F1, an improvement of two to three orders of magnitude.

1. Introduction

Fine grained mushroom classification is a challenging visual recognition task with practical implications for biodiversity monitoring and public safety. Many fungal species differ only in subtle morphological cues such as variations in cap texture, pore patterns, gill structure, or slight shifts in color. These differences are often difficult to distinguish even for expert mycologists. Visual appearance is further affected by illumination, viewpoint, growth stage, and background clutter, which makes reliable identification from a single photograph particularly demanding. At the same time, large annotated fungal datasets remain scarce and costly to curate. This scarcity limits the effectiveness of fully supervised approaches and motivates techniques that can generalize to fine grained categories with limited annotated data. Convolutional neural networks have achieved strong performance on curated mushroom datasets, but they rely on extensive class specific supervision and often struggle to generalize beyond the training distribution. More advanced approaches such as synthetic data generation, three dimensional reconstruction, or metadata driven classification require inputs that are not available for most fungal photographs in natural conditions. Vision and language models such as CLIP offer a compelling alternative because they align images with natural language and support recognition in zero shot or few shot settings. However, it remains unclear whether these models can reliably discriminate highly fine grained biological categories, especially when species share many visual features and differ only subtly. Addressing this uncertainty is central to understanding the usefulness of vision and language models in ecological applications. In this work we investigate whether vision and language models can serve as effective backbones for large scale mushroom classification. Using a balanced dataset of 169 species, we evaluate CLIP based encoders under several adaptation strategies. We begin with strict zero shot evaluation, then introduce prompt enrichment through expert templates, image derived visual attributes, and cleaned common names. We further explore few shot learning using prototype based and linear classifiers, and we study the effect of increasing backbone capac-

ity. Finally, we introduce α-fusion, a simple yet effective mechanism that adjusts the relative contribution of visual and textual features. We study both a global coefficient and species-specific variants, and we show that *species-specific* mixing coefficients α_c combined with a strong OpenCLIP bigG backbone achieve 95.49 percent macro-F1, representing a substantial improvement over all previous configurations. Our results show that language plays a central role when adapting CLIP to visually similar species. Enriched prompt sets consistently outperform simple class name prompts, and cleaned attribute based prompts provide the largest gains in zero shot accuracy. Few shot learning significantly improves performance, particularly when combined with stronger backbones. Lightweight adaptation through prompt design and α-based visual–textual mixing on larger OpenCLIP models proves more effective in our setup than LoRa tuning of a smaller ViT-B/32 backbone. Fusion with ResNet features helps only in restricted conditions and often collapses to the stronger model unless the training distribution is carefully controlled. These observations highlight the importance of linguistic precision, backbone selection, and tuned fusion strategies when applying vision and language models to ecological classification. This paper makes three contributions. First, we provide a unified, reproducible evaluation pipeline for fine-grained mushroom recognition, including zero-shot, prompt-enriched, and few-shot configurations. Second, we present an extensive empirical study that examines prompt engineering, backbone scaling, adapter-based tuning, and visual-textual fusion strategies across 169 species, identifying the combinations that yield the greatest improvements. Third, we introduce a per-class analysis framework that reveals systematic failure modes, including species complexes and morphology-driven ambiguities, and we show how the proposed adaptations mitigate these confusions. Together, these contributions demonstrate that vision and language models can be effectively adapted to highly fine-grained ecological domains when combined with appropriate linguistic and algorithmic design.

2. Related Work

Mushroom Image Classification Early work on mushroom recognition relies on convolutional neural networks trained on small, domain specific datasets. Models such as EfficientNet and Inception have been used for natural habitat imagery, mobile applications, and food safety scenarios. Research in agricultural robotics has explored three dimensional reconstruction, synthetic scene generation, and pose estimation for mushroom detection. These systems perform well in restricted settings but require substantial supervision and do not scale to the large number of visually similar species encountered in fine grained fungal classification.

Fine Grained Visual Categorization Tasks such as bird, insect, or flower identification share many of the difficulties present in fungal classification. The combination of subtle inter class differences, high intra class variability, and limited labeled data has motivated approaches based on transfer learning, strong data augmentation, and localized feature extraction. Surveys in few shot fine grained classification emphasize the importance of powerful pretrained representations that can provide discriminative features even when only a few labeled examples are available.

Vision and Language Models for Biodiversity and Ecology Vision and language models such as CLIP and BLIP provide a natural framework for ecological classification because they align images with textual descriptions and support zero shot evaluation. Prior studies have used CLIP for herbarium specimen retrieval, species classification, and multimodal integration with metadata such as geolocation. While these works show that language guidance can be beneficial, performance on visually dense and highly fine grained categories remains uncertain, particularly when textual descriptions are ambiguous or inconsistent.

Prompt Engineering and Multimodal Integration Prompt design has emerged as a crucial factor in adapting vision and language models to specialized domains. Prompts that encode morphological attributes, habitat descriptors, or common names can improve zero shot accuracy. Several works also explore multimodal integration methods such as prototype based fusion, weighted prediction fusion, and logistic stacking to combine CLIP with supervised vision encoders. These approaches can exploit complementary information, although naive fusion tends to degrade performance when model confidences are misaligned.

Few shot and Lightweight Adaptation Few shot learning methods such as prototypical classifiers and linear probes leverage pretrained embeddings to achieve strong performance with limited supervision. Parameter efficient adaptation techniques, including LoRa and related low rank update methods, further allow targeted tuning of large pretrained models while keeping most parameters fixed. These strategies are particularly valuable in biological classification settings, where comprehensive annotated datasets are difficult to obtain.

3. Method

3.1. Dataset and Preprocessing

We evaluate all methods on a balanced mushroom dataset containing 169 species with approximately 4,080 training images per class, yielding 689,520 total training samples. The validation and test splits contain 15,616 and 15,614 images respectively. All photographs depict single specimens captured under natural conditions, with substantial variation in illumination, viewpoint, background and growth stage.

The dataset is highly fine grained. Many species differ only in subtle variations of cap texture, pore structure, stem morphology or color gradients, and several belong to morphological complexes where genetic distinctions do not fully manifest visually.

We use the preprocessing transforms supplied by OpenCLIP for the OpenCLIP backbones, and analogous resize–crop–normalize pipelines for the HuggingFace CLIP model used in the LoRA experiments. Images are loaded with PIL, converted to RGB, resized, center cropped and normalized. Species names are standardized through a vocabulary file, and metadata files include BLIP generated captions, cleaned captions and attribute descriptors that support prompt construction.

For efficiency, all image embeddings are precomputed. Using batches of 64 images per forward pass, CLIP encodes each image with $model.encode\_image$, and all embeddings are L2-normalized so that cosine similarity corresponds to a dot product. Embeddings, labels and relative paths are stored in compressed NPZ files and reused across all experiments.

3.2. Baseline Architectures

We compare two baselines that motivate the need for adaptation.

ResNet supervised baseline. A ResNet classifier trained on all labeled images provides a fully supervised reference point that represents standard convolutional performance on this dataset.

Zero shot CLIP baseline. CLIP predicts species by comparing image embeddings to text embeddings of class name prompts. Zero shot classification is implemented as a normalized cosine similarity between image features and text features. Across CLIP backbones, zero shot performance remains limited due to the extreme fine grained nature of the dataset, motivating the use of enriched prompts and adaptation strategies.

3.3. Prompt Generation

Prompt design is known to play a critical role in adapting vision-language models, such as CLIP, to downstream tasks [11]. In line with prior work on hand-crafted templates, prompt ensembling, and prompt learning, we construct four families of prompts that differ in how they inject morphological knowledge, visual attributes, and lexical variants into the label space. All prompts are encoded with the CLIP text encoder, ℓ_2-normalized, and aggregated into class-level text prototypes that are reused across zero-shot and few-shot experiments.

Prompt Set 1: Template-Driven Prompts with Knowledge-Base Attributes. A common strategy for CLIP adaptation is to insert class names into generic natural language templates and ensemble their embeddings. We follow this practice but specialize the templates to the mycological domain. Specifically, we use concise image templates such as *"a photo of* $\{CLASS\}$*"* and *"a close-up of* $\{CLASS\}$ *mushroom"*, and augment them with curated morphological descriptors from an expert knowledge base (e.g., cap color and texture, hymenophore type, stem characteristics, ecological context). Example prompts include *"Amanita muscaria growing naturally in birch and pine forests"*, *"Amanita muscaria showing white stem with ring"*, and *"pristine Chlorociboria aeruginascens specimen for study"*. This design mirrors template-based prompt engineering while explicitly encoding fine-grained morphological cues that are often underrepresented in generic language.

Prompt Set 2: Image-Derived Descriptors. Recent work has explored using image-conditioned text generation to strengthen visual–textual alignment [7, 8]. Motivated by this, we derive visually grounded prompts from the dataset itself. We run BLIP [7] to obtain multiple captions per image and parse them with mushroom-specific lexicons to extract attributes for color (e.g., "red-capped", "tan-colored"), surface texture (e.g., "rough cap texture", "smooth cap surface"), shape and growth form (e.g., "pear-shaped puffball", "bulbous base", "overlapping brackets"), habitat or season (e.g., "growing in grassy areas", "on wood chips", "late summer to fall"), and photographic style (e.g., "macro close-up", "field guide photo", "naturally lit"). Attributes are aggregated per image by field-wise majority vote and then summarized per species by counting frequencies and retaining the dominant values. We insert the resulting descriptors into short templates to form attribute-only and enriched prompts. This procedure can be viewed as a lightweight, attribute-focused instance of automatic prompt construction driven by image-derived evidence.

Prompt Set 3: Common and Latin Names. Several works have highlighted that the lexical form of labels (including synonyms, paraphrases, and alternative names) can significantly affect CLIP performance [10, 11]. In ecological domains, species are frequently referred to by both Latin binomials and common names, and the latter may be better represented in web-scale corpora. We therefore expand the label space by collecting English common names and scientific names from Wikidata and Wikipedia, followed by automated filtering to remove non-English or ambiguous variants and to normalize formatting. For example, prompts containing the common name "fly agaric" complement those using the Latin name *Amanita muscaria*. This label-space expansion aims to better match CLIP's pre-training distribution and to reduce mismatches between dataset labels and natural usage.

Prompt Set 4: Combined Prompt Ensemble. Following the practice of prompt ensembling in CLIP [11] and subsequent work, we construct a combined family that merges Sets 1-3 into a unified prompt pool per species. We study two aggregation strategies: (i) *naïve pooling*, where all prompt embeddings for a class are averaged into a single text prototype; and (ii) *per-prompt pooling*, which retains individual embeddings and aggregates them only after computing image–text similarities. The latter is analogous in spirit to ensemble inference over prompts, preserving diversity across templates, attributes, and lexical variants. In our experiments, per-prompt pooling yields more robust performance, particularly in low-shot regimes.

Embedding Construction and Capping. For each family, we deduplicate near-identical prompts, standardize species naming, and cap the number of prompts per class to maintain a balanced and computationally tractable prompt set. All prompt embeddings are precomputed, stored, and reused for zero-shot classification, few-shot adaptation, and fusion experiments described in subsequent sections.

3.4. Customization Approaches

To modify CLIP to this fine-grained domain, we evaluate three complementary strategies.

Backbone scaling. We evaluate diffrent OpenCLIP backbones of increasing capacity, including ViT-B/32 quickgelu, ViT-L/14, and the larger bigG model. Larger encoders produce higher dimensional and more expressive features that capture subtle morphological differences, leading to significant gains but longer runtime with higher memory requirments.

Few-shot learning. We study modification with k labeled images per class using two standard methods: prototype classification and linear probing on frozen CLIP embeddings. For each backbone, we first precompute and cache ℓ_2-normalized image embeddings for the train, validation and test splits. Given a shot value k, we sample k support examples per class from the training split (with replacement if necessary), yielding a support set $\{(x_i, y_i)\}$ that covers all 169 species.

The prototype classifier represents each class by the mean of its support embeddings in the CLIP space, followed by normalization. Predictions are obtained by cosine similarity to these class prototypes. The linear probe instead trains a single fully connected layer on top of the frozen embeddings using cross-entropy loss. We optimize the linear layer with AdamW for up to 200 epochs and keep the checkpoint with the lowest validation loss (when validation embeddings are available), using standard hyperparameters such as the learning rate or *lr* (the rate of optimization change), and weight decay or *wd* (regularisation of model weights).

We additionally consider prompt-aware variants of both methods. For the prototype classifier, we combine the image prototype for class c with its text embedding $\mathbf{t}_c$ derived from the prompt families, forming

$$\tilde{\mathbf{p}}_c = \mathrm{norm}\big(\alpha\, \mathbf{p}_c + (1-\alpha)\, \mathbf{t}_c\big),$$

where $\alpha \in [0,1]$ controls the relative contribution of visual and textual cues. For the linear probe, we construct mixed support features by adding the class-specific text embedding to each support vector before normalization, again with a controllable mixing coefficient α.

In both cases, we sweep over a large grid of hyperparameters α, lr, wd and select the best lr and wd according to the balanced account values of the validation accuracy. This scalar mixing captures the varying utility of linguistic information across backbones and data regimes.

We then extend this idea to *species-specific* mixing coefficients α_c, which allow different classes to place different weight on visual versus textual cues. The per-class fusion operates on the same mixed prototypes as in the global case but replaces the scalar α with α_c.

LoRA adapter fine-tuning. To adapt CLIP in a parameter-efficient way, we insert LoRA adapters into the last two transformer blocks of both the text and vision encoders of the `Vit-B-32-quickgelu` model pretained on the `Laion400M-e32` dataset. Concretely, we add rank-$r = 128$ low-rank update matrices with scaling factor $\alpha = 256$ and dropout 0.1 to the query, key, value and output projections of the self-attention layers in blocks 10 and 11 of each encoder. All non-adapter CLIP parameters are frozen; only the LoRA parameters are trainable, so more than 99% of the model remains fixed.

Training is performed on a large subset of the mushroom training split (up to 200 000 images, sampled in a class-balanced way) with enhanced data augmentation. The vision encoder receives strong geometric and photometric transformations (random resized crops, flips, rotations, affine transforms, color jitter, blur, grayscale, random erasing) followed by CLIP-style normalization, while the validation set uses a standard resize–center-crop pipeline.

For the classifier, we precompute a text prototype for each species by encoding a small set of cleaned prompts (including common names) with the CLIP text encoder and averaging the ℓ_2-normalized text features. During training, each image is mapped to a normalized image embedding and scored against the fixed matrix of text prototypes via cosine similarity. We optimize a temperature-scaled, label-smoothed cross-entropy with difficulty-aware weighting and weight decay on the LoRA parameters via AdamW. The learning rate follows a warm-up (three epochs) plus cosine decay schedule, with additional adjustments based on the current loss and gradient clipping. We train for up to 200 epochs with early stopping after 50 epochs without im-

provement on the validation top-1 accuracy, and select the best checkpoint according to this metric.

3.5. ResNet and CLIP Feature Fusion

To explore whether supervised visual features complement CLIP embeddings, we evaluate three fusion strategies that combine predictions from a ResNet classifier and CLIP.

Prototype fusion. Each model computes class prototypes, and predictions are made based on combined similarity scores.

Alpha fusion. CLIP and ResNet probability distributions are combined with a weighted average tuned on validation data.

Logistic regression stacking. A shallow classifier is trained on concatenated prediction vectors to exploit complementary confidence patterns. These methods test whether supervised and multimodal representations provide complementary information. In practice, fusion performance depends strongly on calibration and often defaults to the stronger model.

4. Experiments

4.1. Zero-Shot Evaluation

We begin by evaluating CLIP in a strict zero-shot setting using several prompt families. Image embeddings are precomputed with the CLIP vision encoder and ℓ_2-normalized. Text embeddings for each class are obtained by encoding all prompts associated with that class, normalizing each prompt embedding, mean-pooling them and renormalizing. Zero-shot classification is implemented as a cosine similarity between normalized image features and the class-level text embedding matrix.

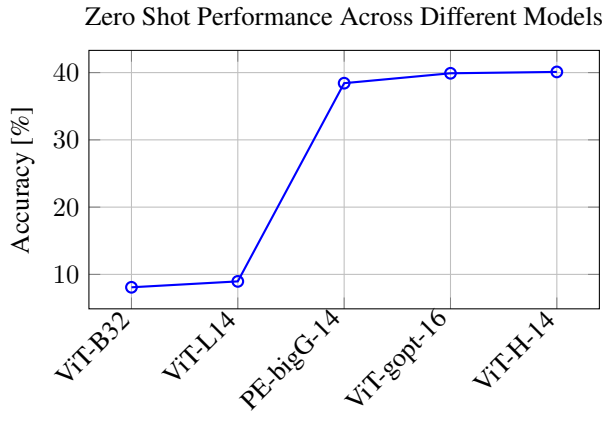


Figure 1. zero-shot performance across different CLIP models using image encoding and class labels encoding. The plot shows that the more complex the model is, the better its performance, peaking at 40.10% macro F1 accuracy.

Model	Setting	Top-1(%)↑	Top-5(%)↑	Bal. Acc.(%)↑	**Macro-F1(%)↑**
ResNet	fully supervised	–	–	–	**80.80**
ResNet	50% images / class	0.77	2.95	0.65	**0.02**
ResNet	50% classes	0.62	2.85	0.59	**0.01**
CLIP ViT-B/32	50% images / class	10.15	29.61	10.15	**7.10**
CLIP ViT-B/32	50% classes	10.15	29.61	10.15	**7.10**
CLIP ViT-B/32	zero-shot	21.26	40.55	12.56	**8.08**
CLIP bigG	zero-shot	53.50	85.84	44.64	**38.43**

Table 1. Baseline ResNet and zero-shot CLIP performance on the held-out split. All numbers are percentages. Bal. Acc. denotes balanced accuracy.

Using only class-name prompts the best ViT-B/32 configuration reaches 21.3% top-1 accuracy, 40.6% top-5 accuracy, 12.6% balanced accuracy and 8.1% macro-F1 on the validation set. Attribute-based and enriched prompt families obtain very similar performance, with differences typically within one to two percentage points across all metrics. However, as can be observed with Figure 1, With the larger backbones, performance increases significantly.

Nevertheless, the enriched prompt families are useful for analysis and downstream adaptation: they structure class descriptions into interpretable morphological cues (cap colour, texture, hymenophore type, habitat, lighting) and provide a principled way to integrate BLIP-derived information into CLIP's text encoder. In later sections we use these enriched descriptions as a basis for few-shot learning and prompt-aware fusion.

4.2. Few-Shot Evaluation

We next evaluate few-shot adaptation with $k \in \{1, 5, 10, 20, 50, 100\}$ shots (labeled images) per class. For each shot value, we sample k support examples per species from the training split using a fixed (42) random seed, ensuring that every class contributes the same number of labeled images. All methods operate on frozen CLIP embeddings, which are precomputed and ℓ_2-normalized for the train, validation and test splits.

Next, we performed a large hyperparameter sweep for all shots to determine the best combination, considering 6 different backbones (listed in Figure 1), prompt types, and various variables, including $\alpha \in [0, 1]$, learning rates in $[1 \times 10^{-3}, 3 \times 10^{-1}]$, and weight decays in $[0, 5 \times 10^{-4}]$, with two few shots methods, prototyping and linear probing.

As can be observed in the figure 2, performance improves rapidly as more shots become available, and larger CLIP backbones consistently outperform smaller ones. Making the best few-shot configurations to include models using the combined prompts[3.3] (which is consistent with what had been observed in the prompt analysis), 100 image samples per class, learning rate of 3×10^{-1}, and a weight decay of 0 for the ViT-B/32 backbone and 1×10^{-4} for the PE-Core-bigG-14-448 backbone.

However, this performance can be further improved by

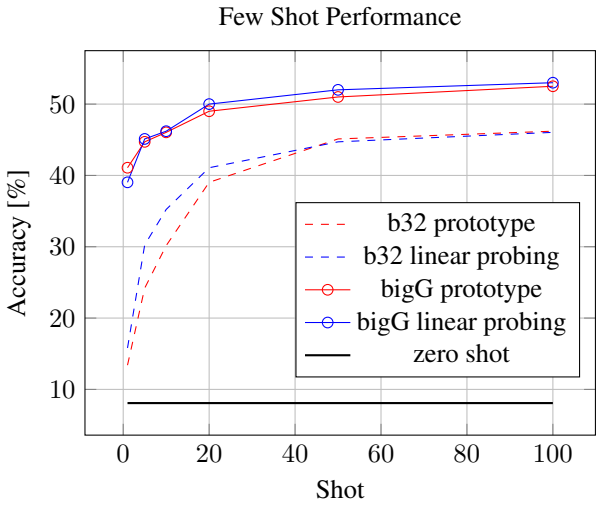


Figure 2. Few-shot performance for B32 and bigG models with prototypes and linear probing. Zero-shot baseline is shown as a dotted black line. We can observe that the few-shot performs significantly better, and larger models and linear probing both perform better than their counterparts.

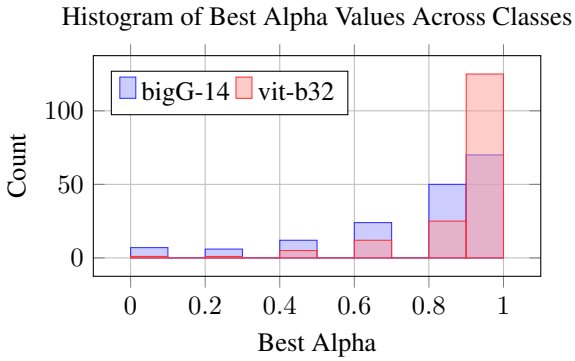
moving from global alpha parameters to per-class analysis. As can be observed in figure 3, the best accuracy per class has different alpha values. Thus, if we restrict ourselves to a single global value, we degrade our models. In fact, implementing per-class alpha increased macroF1 accuracy to 78% for the Vit-b32 model and 95% for the bigG-14 backbone.

Thus, we observed that the large model already behaves so much better than the fully supervised ResNet model (85.1%), even though it requires only 100 examples per class. However, while larger mode proved to perform significantly better. They also consume significantly more resources, making them unmanageable in many scenarios. Thus, when exploring different CLIP customization approaches, the standard backbone was Vit-b32, which peaks at 78%, which is only slightly worse than resNet. Thus, requiring an investigation into other customizations, such as CLIP adapters.

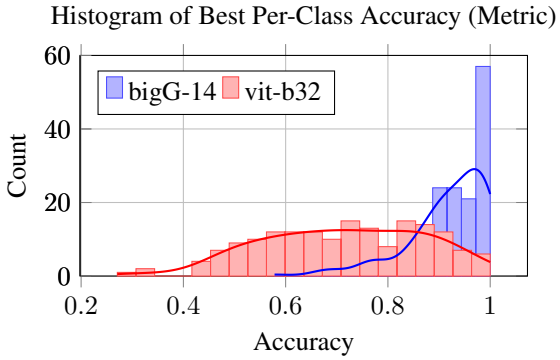
4.3. Adapter-Based Adaptation

We compare the parameter-efficient LoRA fine-tuning described in Section 3 with few-shot linear probing on frozen CLIP embeddings. The backbone used for LoRA is shown in Table 2, and we additionally explore whether varying the LoRA rank can further improve accuracy (Figure 4).Among five rank settings evaluated, the model achieves peak accuracy at rank 128. Balancing accuracy with computational efficiency, we select rank 128 as the optimal configuration.

The best checkpoint (LoRA on the ViT-B/32 backbone with openai pretraining) achieves 88.8% top-1 accuracy,



(a) Histogram of best per-class alpha distribution. Showing that the bigG model relies on prompts more compared to the small B32 model.



(b) Histogram and Kernel Density estimation of per-class accuracy. Here we see that the larger backbone performs significantly better, with the density curve being narrower and centred closer to the 1.0 value. Implying more classes hit 100% accuracy.

Figure 3. Comparison of per-class performance for the best b32 linear probing model, vs best model overall (bigG). Specifically best alpha distribution per class, and the corresponding per-class accuracies. Demonstrating a direct relation between the size of backbone, and best per-class accuracy.

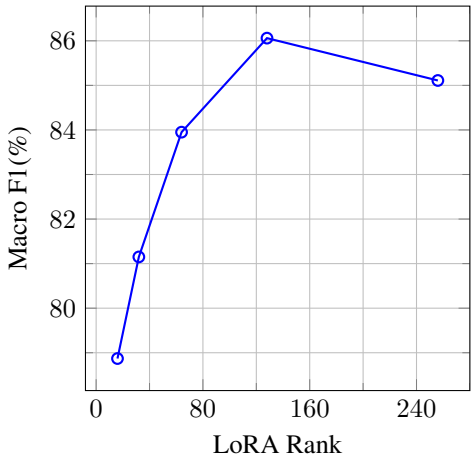


Figure 4. Performance across different LoRA ranks.

97.4% top-5 accuracy, 86.9% balanced accuracy and 86.4% macro-F1 on the test set, substantially improving over zero-shot CLIP and matching or surpassing a fully supervised ResNet baseline with a similar parameter count.

These results show that the adapter performs significantly better than few-shots, when ran with the same backbone. which, considering the improved performance with a larger backbone seen in Section 4.2, implies that an adapter on the same setup would have results which significantly outperform the baseline resNet. However, due to constraints in computational resources and time, we encountered memory and runtime limitations that precluded a comprehensive exploration of this direction.

4.4. Fusion Experiments

To test whether supervised visual features complement CLIP embeddings, we evaluate three fusion strategies that combine predictions from a ResNet classifier and CLIP. Prototype fusion computes separate class prototypes for ResNet and CLIP and scores test images by summing similarity scores from both models. Alpha fusion combines prediction distributions with a weight tuned on validation data. Logistic-regression stacking trains a shallow classifier on concatenated prediction vectors from both models. Fusion performance depends strongly on calibration. When the supervised ResNet classifier provides meaningful confidence estimates, fusion provides modest improvements. However, when ResNet predictions are poorly calibrated, alpha sweep experiments select $\alpha = 0.0$, showing that fusion collapses to the stronger CLIP predictor. These results are consistent with prior observations that multimodal ensembles require balanced and reliable confidence signals to be effective.

4.5. Final Model Performance

Our best-performing configuration combines prompt enrichment, few-shot linear probing, backbone scaling, and species-specific visual-textual fusion. Using the OpenCLIP PE-Core-bigG-14-448 backbone with 100 shots per class, a linear probe trained on frozen image embeddings, and per-class mixing coefficients α_c estimated on the validation set, the model achieves:

- Top-1 accuracy: 95.84 percent
- Top-5 accuracy: 99.72 percent
- Balanced accuracy: 96.25 percent
- Macro-F1: 95.49 percent

This represents a substantial improvement over zero-shot performance and surpasses supervised baselines. The combination of linguistic enrichment, large-scale vision encoders and species-specific weighting of image and text cues is crucial for fine-grained mushroom identification.

Model	Setting	Top-1(%)↑	Top-5(%)↑	Bal. Acc.(%)↑	**Macro-F1(%)↑**
LoRA ViT-B/32 (HF)	adapter, patch32	86.92	98.21	84.93	**83.95**
LoRA ViT-B/32	adapter, laion400m_e32	87.57	97.28	85.37	**84.83**
LoRA ViT-B/32	adapter, openai	88.84	97.35	86.85	**86.36**
CLIP ViT-B/32	few-shot, $\alpha = 1.0$, 100 shots	70.10	92.28	69.02	**65.77**
CLIP ViT-B/32	few-shot, α_c, 100 shots	74.42	76.76	94.80	**78.08**
CLIP bigG	few-shot, $\alpha = 1.0$, 100 shots	89.83	98.86	90.29	**88.58**
CLIP bigG	few-shot, α_c, 100 shots	95.84	99.72	96.25	**95.49**

Table 2. Adapter-based and few-shot CLIP results. "Shots" denotes labeled images per class; α and α_c are global and species-specific visual-textual mixing coefficients. All numbers are percentages.

4.6. Error Analysis

To diagnose the remaining failure cases, we perform a structured error analysis on the LoRA-adapted CLIP model using the validation predictions collected during training. For each image, we record (i) the top-1 prediction, (ii) the top-5 predictions, and (iii) the true label. Species aggregate misclassified samples, and we additionally count the most frequent confusion pairs (Table 3). We observe that the majority of errors arise from fine-grained species complexes where morphological boundaries are subtle or controversial ¯[5].

Table 3. Most common confusion pairs in the model predictions.

True label	**Predicted as**	**Count**
Parmelia sulcata	Hypogymnia physodes	6735
Fomes fomentarius	Phellinus igniarius	4095
Fomes fomentarius	Fomitopsis pinicola	4027
Fomes fomentarius	Fomitopsis betulina	3786
Leccinum aurantiacum	Leccinum versipelle	3626
Fomitopsis pinicola	Fomitopsis mounceae	3457
Xanthoria parietina	Vulpicida pinastri	3410
Amanita muscaria	Amanita persicina	2976
Pleurotus pulmonarius	Pleurotus ostreatus	2879
Pleurotus ostreatus	Pleurotus pulmonarius	2820
Fomitopsis pinicola	Fomes fomentarius	2743

Examples include pairs such as *Fomitopsis pinicola* vs. *Fomitopsis mounceae*[5], *Pleurotus pulmonarius* vs. *Pleurotus ostreatus*[14], and several species within the *Cortinarius* genus. These pairs also dominate the confusion matrix generated by our ErrorAnalyzer module, confirming that residual errors are largely biological rather than algorithmic.

Top-5 errors reveal similar patterns: most misses occur within the same genus, and the predicted top-5 set often contains visually plausible alternatives. This suggests that single-view image classification is inherently ambiguous for some taxa, even when using enriched prompts and large backbones. Such cases are likely resolvable only with additional metadata (habitat, color changes by season, substrate) or multi-view images.

Overall, the error statistics are consistent with the quantitative trends: most confusions occur among closely re-

lated species, and the remaining mistakes largely align with known taxonomic ambiguities, reflecting intrinsic limitations of the visual modality for certain mushroom species.

5. Conclusion

This work investigates the use of vision and language models for large scale, fine grained mushroom classification. Using a balanced dataset of 169 species, we evaluate zero shot CLIP performa nce, introduce multiple forms of prompt enrichment, and study few shot adaptation across a range of backbones. Our results show that linguistic information plays a central role in adapting CLIP to visually similar species. Prompt enrichment consistently improves zero-shot accuracy, and cleaned attribute-based prompts provide the largest gains. Few-shot learning offers a powerful mechanism for further adaptation and benefits substantially from increased model capacity. Scaling to larger OpenCLIP backbones yields significant improvements, and our final method, which combines prompt enrichment with a species-specific visual-textual mixing scheme, achieves 95.49% Macro-F1. These findings lead to several broader insights. First, prompts matter: carefully constructed textual descriptions help disambiguate species that differ only in subtle morphological cues. Second, few-shot learning is highly effective in fine-grained domains where a small number of labeled examples can greatly improve discriminative power. Third, backbone scaling is critical, since larger vision encoders capture fine structures that smaller models fail to resolve. Finally, visual–textual fusion is essential for optimal performance: even a single global α provides substantial gains, and our per-species analyses reveal that different taxa rely on visual or linguistic cues to varying degrees. There are several promising directions for future work. One avenue is to apply parameter efficient adapters such as LoRa to larger backbones, which may combine the benefits of model capacity with efficient fine tuning. Another is to revisit multimodal fusion with stronger supervised vision architectures that may offer more complementary signals. Integrating ecological metadata such as habitat, season or location could help resolve ambiguities within species complexes. Finally, synthetic data generation and generative augmentation may further improve robustness in scenarios with limited or biased real world imagery.

References

[1] Rab Nawaz Bashir, Olfa Mzoughi, Nazish Riaz, Muhammed Mujahid, Muhammad Faheem, Muhammad Tausif, and Amjad Rehman Khan. Mushroom species classification in natural habitats using CNNs. *IEEE Access*, 2024.

[2] Renjun Cai. Automating bird species classification: A deep learning approach with CNNs. *Journal of Physics: Conference Series*, 2664(1):012007, 2023.

[3] Gulce Berfin Ercan, Melis Baran, Ecem Konca, Ilhan Mert Cetin, and Ilker Korkmaz. Mushroom classification using machine learning. In *International Conference on ICT Innovations*, pages 159–173. Springer, 2024.

[4] Muhammad Waleed Gondal, Jochen Gast, Inigo Alonso Ruiz, Richard Droste, Tommaso Macri, Suren Kumar, and Luitpold Staudigl. Domain aligned CLIP for few-shot classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5721–5730, 2024.

[5] John-Erich Haight, Karen Nakasone, Gary Laursen, Scott Redhead, D. Lee Taylor, and Jessie Glaeser. Fomitopsis mounceae and f. schrenkii-two new species from north america in the f. pinicola complex. *Mycologia*, 111:1–19, 2019. 7

[6] Pratham Kaushik and Savinder Kaur. Deep learning-based mushroom species classification: Analyzing the performance of CNN on diverse fungi. In *2024 13th International Conference on System Modeling & Advancement in Research Trends (SMART)*, pages 21–27. IEEE, 2024.

[7] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. *CoRR*, abs/2201.12086, 2022. 3

[8] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. 3

[9] Lei Liu, Linzhe Yang, Feng Yang, Feixiang Chen, and Fu Xu. CLIP-driven few-shot species-recognition method for integrating geographic information. *Remote Sensing*, 16(12): 2238, 2024.

[10] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. In *International Conference on Learning Representations (ICLR)*, 2023. 3

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 3, 4

[12] Jie Ren, Changmiao Li, Yaohui An, Weichuan Zhang, and Changming Sun. Few-shot fine-grained image classification: A comprehensive review. *AI*, 5(1):405–425, 2024.

[13] Maya Sahraoui, Youcef Sklab, Marc Pignal, Régine Vignes Lebbe, and Vincent Guigue. Leveraging multimodality for biodiversity data: Exploring joint representations of species descriptions and specimen images using CLIP. *Biodiversity Information Science and Standards*, 7:e112666, 2023.

[14] A. V. Shnyreva and O. V. Shtaer. Differentiation of closely related oyster fungi Pleurotus pulmonarius and P. ostreatus by mating and molecular markers. *Russian Journal of Genetics*, 42(5):539–545, 2006. 7

[15] Tuan-Anh Yang and Minh-Quang Nguyen. Mushroom for improvement: Prototypical few-shot learning with multimodal fungal features. *arXiv preprint arXiv:2501.0819*, 2025.

Appendix

A. Results Expanded

Table 4 summarizes the full set of experimental results across all evaluated models, supervision regimes, and adaptation strategies. The table reports performance for fully supervised ResNet baselines, reduced-supervision ResNet configurations (50% images and 50% classes), zero-shot CLIP models, parameter-efficient LoRA adapters, and few-shot CLIP adaptation on multiple backbones. Fully supervised ResNet performance improves with model depth but deteriorates sharply when supervision is reduced. In particular, ResNet models trained on half the images or half the classes exhibit near-zero macro-F1, highlighting the limited ability of closed-set supervised CNNs to generalize to unseen data. In contrast, CLIP maintains substantially higher performance under limited supervision and demonstrates strong zero-shot capability, with accuracy improving significantly as backbone capacity increases. Parameter-efficient LoRA adaptation yields large gains on smaller CLIP backbones, outperforming supervised ResNet baselines while updating only a small fraction of model parameters. Few-shot adaptation further improves performance, especially on the bigG backbone, where combining prompt enrichment with visual-textual mixing achieves the best overall result of 95.49% macro-F1.

Model	**Subtype**	**Top-1 (%)**	**Top-5 (%)**	**Balanced Acc. (%)**	**Macro-F1 (%)**
ResNet	Fully supervised (18)	–	–	–	80.80
ResNet	Fully supervised (152)	–	–	–	85.10
ResNet	50% images	0.77	2.95	0.65	0.016
ResNet	50% classes	0.62	2.85	0.59	0.007
CLIP	50% images	10.15	29.61	10.15	7.10
CLIP	50% classes	10.15	29.61	10.15	7.10
Zero-shot	B-32	21.26	40.55	12.56	8.08
Zero-shot	bigG	53.50	85.84	44.64	38.43
LoRA Adapter	ViT-B/32 (HF)	86.92	98.21	84.93	83.95
LoRA Adapter	ViT-B/32 (LAION)	87.57	97.28	85.37	84.83
LoRA Adapter	ViT-B/32 (OpenAI)	88.84	97.35	86.85	86.36
Few-shot	B-32 (univ. α)	70.10	92.28	69.02	65.78
Few-shot	B-32 (100 shots)	74.42	76.76	94.80	78.08
Few-shot	bigG (univ. α)	89.83	98.86	90.29	88.58
Few-shot	bigG (100 shots)	95.84	99.72	96.25	95.49

Table 4. Overarching table demonstrating the results of our analysis in its entirety, including the model variation and its performance assessed as top-1, top-5, balanced accuracy, and macro-F1. The table shows that adapter-based CLIP models outperform supervised ResNet baselines under limited supervision, and that larger CLIP backbones yield substantial gains in fine-grained classification accuracy.

B. AI usage declaration

In this project, we used generative AI tools as *aids*. All decisions about the design of the experiments, the implementation of the code and the interpretation of the results were made by us.
We used the following tool:

- **ChatGPT (OpenAI, GPT-5.1, 2025):** Used as a writing assistant to improve grammar, clarity and structure of paragraphs that we had written ourselves, and as a summarization aid for our own notes and intermediate results.

We did *not* use generative AI to generate figures or experimental results. We bear full responsibility for all content of this report, including any sections in which we used AI-assisted wording.