Mushroom for Improvement: Prototypical Few-Shot Learning with Multimodal Fungal Features

Notebook for the LifeCLEF Lab at CLEF 2025

Tuan-Anh Yang[1,2], Minh-Quang, Nguyen[1,2]

[1]VNU-HCM University of Science, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam

Abstract

We present a multimodal few-shot classification pipeline for the FungiCLEF 2025 challenge, addressing the task of fine-grained fungal species recognition from sparse and heterogeneous observations. Our approach integrates visual features from three pretrained image encoders—BioCLIP, SigLIP ViT-B/16, and DINOv2 - with textual descriptions and structured metadata using a unified multimodal embedding. The model is trained in two stages: initial supervised pretraining of a multimodal encoder followed by prototypical network fine-tuning under an episodic few-shot regime. We further apply an observation-level reranking strategy that aggregates predictions across multiple images per observation via a weighted voting scheme. Evaluation on the official FungiCLEF 2025 public and private test sets demonstrates strong performance, with Recall@5 scores of 0.57079 and 0.55498, respectively. Ablation results confirm the additive benefit of combining image, text, and metadata features. The code is available at https://github.com/YangTuanAnh/FungiCLEF2025

Keywords

fine-grained visual categorization, few-shot learning, prototypical networks, multimodal representation learning, foundational models, CEUR-WS

1. Introduction

The FungiCLEF 2025 [1] Challenge addresses few-shot recognition of fungi species using real-world data comprising multiple photographs, rich metadata (e.g., location, substrate, toxicity), satellite imagery, and meteorological variables. The task requires models to return a ranked list of species predictions per observation, despite the challenges of large class diversity and many rare or under-recorded species with limited training data.

The motivation behind this challenge lies in the need to support mycologists, citizen scientists, and nature enthusiasts in species identification while contributing to biodiversity data collection. To be practical for large-scale citizen science projects, models must efficiently handle numerous classes—including those with scarce observations—and operate under limited computational resources. Importantly, rare species are often excluded from training data, complicating AI models' ability to recognize them.

In this work, we introduce:

- Introduce a robust multimodal classification framework integrating image, text, and metadata features for fungi species recognition.
- Propose a two-stage training strategy combining supervised encoder pretraining with prototypical network fine-tuning for improved few-shot learning performance.
- Demonstrate the effectiveness of the method on the FungiCLEF 2025 challenge dataset, achieving significant gains in few-shot classification accuracy.

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain
ytanh21@apcs.fitus.edu.vn (T. Yang); nmquang21@apcs.fitus.edu.vn (M. Nguyen)
https://yangtuananh.dev/ (T. Yang); https://htamlive.github.io/quangmnguyen.github.io/ (M. Nguyen)
0009-0005-3140-8046 (T. Yang); 0009-0008-3520-4624 (M. Nguyen)

2. Related Work

Fine-grained visual categorization (FGVC) focuses on distinguishing between visually similar categories, such as species or subspecies, and often requires models to learn subtle appearance differences under limited supervision [2, 3]. Few-shot learning methods address the scarcity of labeled data by enabling generalization to novel classes with only a few examples; metric-based approaches like Prototypical Networks [4] learn class prototypes in an embedding space and classify queries based on their distances to these prototypes. Multimodal representation learning has been advanced by models such as CLIP [5] and SigLIP [6], which align image and text embeddings via contrastive objectives, enabling robust cross-modal understanding. These foundational models have been adapted to specialized domains, including biology, through domain-specific fine-tuning [7].

3. Dataset

Figure 1: Obs 1: Close-up of fungi.
Feb, Garden, Dead wood

Figure 2: Obs 500: Piece of mold or fungus.
Nov, Woodland, Bark

Figure 3: Obs 1000: Three mushrooms in grassy area.
May, Grassland, Soil

Figure 4: Sample observations from FungiCLEF 2025 showing image diversity, textual notes, and key metadata (month, habitat, substrate).

We used the official **FungiCLEF 2025** [1] dataset, which uses a sample of the **FungiTastic** dataset [8]. The training and validation sets include fungal observations from the Atlas of Danish Fungi submitted before 2024. Each entry features expert-annotated images and rich metadata, including satellite data, weather, timestamps, locations, substrate, habitat, and toxicity. Most entries are fully annotated.

For the official evaluation, the test set comprises a separate collection of images that remained unpublished until the challenge concluded. While partial test results were made publicly available during the competition, the outcomes on the private subset were revealed after the submission deadline.

Table 1
Statistical overview of dataset splits. Each image is accompanied by tabular metadata and auto-generated text. Each class has between 1–4 observations in the training set.

Split	Observations	Images	Classes
Training	4,293	7,819	2,427
Validation	1,099	2,285	570

4. Method

4.1. Image Encoders

The FungiTastic [8] paper evaluated the performance of BioCLIP [7], CLIP[5], and DINOv2[9] on few-shot image classification tasks, identifying BioCLIP[7] as the strongest baseline due to its domain-

specific training on biological imagery. Building on this finding, we investigate whether combining features from all three models can improve classification performance beyond the baseline.

We extract visual features using three pretrained models. **BioCLIP** [7] is a vision-language model fine-tuned on biological image data, which enhances its ability to generalize across diverse species. **SigLIP ViT-B/16** [6] is a contrastive vision-language model based on the ViT-B/16 architecture and trained with a sigmoid loss, offering improved embedding robustness over traditional CLIP-style objectives [5]. **DINOv2 Base** [9] is a self-supervised vision transformer that captures fine-grained spatial and texture-level representations.

4.2. Text and Metadata Features

Textual descriptions and structured metadata are encoded using both rule-based and statistical methods. For textual descriptions, we extract interpretable features such as color, shape, texture, growth pattern, habitat, and size using regular expression matching. These rule-based attributes are complemented by a TF-IDF vectorizer, which captures general linguistic patterns in the descriptions.

Structured metadata, including categorical fields like month, habitat, and substrate, is encoded using one-hot encoding. A fixed schema—defined based on the training set—ensures consistent encoding across training, validation, and test splits. The text and metadata features are concatenated into a single vector, which is optionally included in the classification model.

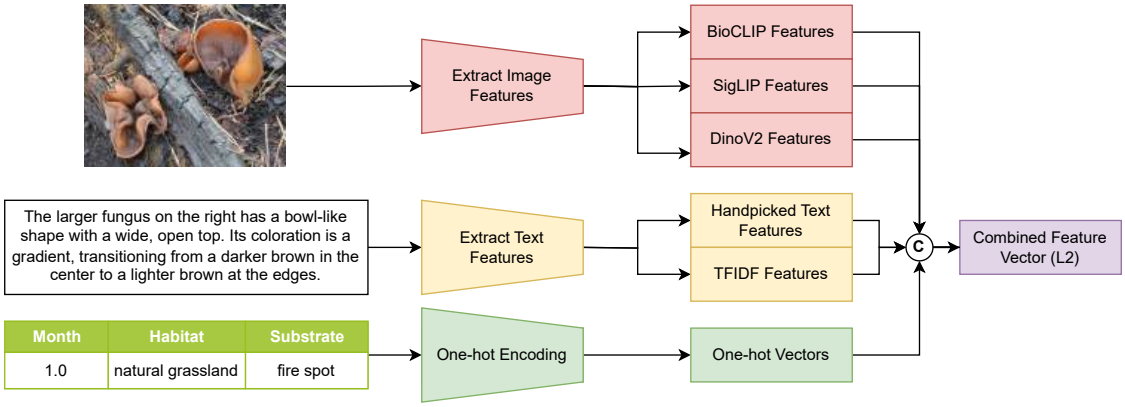
Table 2
Keyword categories and extracted terms used in fungal text feature extraction. Synonyms are mapped to the same class index within each category, proposed by Claude 3.5 Sonnet [10] for fungal descriptors.

Category	Keywords
Color	white, cream, yellow, orange, red, pink, purple, blue, green, brown, black, gray/grey, tan, golden, beige, buff, olive, rusty
Shape	spherical, round, globose, ball, cap, pileus, umbrella, conical, cone, flat, convex, depressed, shelf, bracket, club, coral, fan, cup, disc, bell
Texture	smooth, slimy, sticky, viscid, rough, bumpy, warty, scaly, fibrous, hairy, velvety, fuzzy, ridged, wrinkled, grooved, pitted, powdery, granular, cracked
Growth Pattern	cluster, clustered, group, gregarious, scattered, solitary, single, individual, caespitose, troops, fairy ring, circle, row, line, tuft, dense, packed
Habitat	soil, ground, earth, wood, log, trunk, stump, branch, leaf, leaves, needle, needles, litter, moss, grass, dung, manure, compost, mulch
Mycological Terms	mycelium, hyphae, spore, fruiting body, basidium, basidia, ascus, asci, cystidia, stipe, pileus, lamellae, gills, pores, annulus, volva, universal veil, partial veil, hymenium, cap, stem, stalk

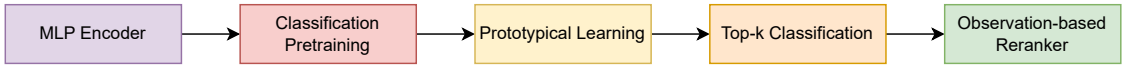
4.3. Multimodal Feature Fusion

To form a unified representation of each fungal observation, we concatenate embeddings from three distinct modalities: image embeddings (from BioCLIP, SigLIP, and DINOv2), structured metadata (encoded as one-hot vectors), and textual features (both rule-based and TF-IDF). The resulting multimodal feature vector is L2-normalized and used as the input to the metric-based few-shot classification pipeline.

Figure 5: Word cloud of common species descriptions from the Fungi dataset. Popular descriptive words such as "texture", "shape", "structure", "color" suggests that the captions include these attributes about the observations.



(a) Multimodal feature extraction pipeline.



(b) Training and prediction pipeline.

Figure 6: Overview of the multimodal processing pipeline. (a) Feature extraction from each modality. (b) Training and prediction based on the extracted features.

4.4. Two-Stage Training Strategy

Our training pipeline follows a two-stage approach. In the first stage, the multimodal encoder is pretrained using standard supervised classification. In the second stage, we adapt the encoder for few-shot learning using a prototypical network [4].

4.4.1. Stage 1: Supervised Encoder Pretraining

The encoder is initially trained in a fully supervised setting with a cross-entropy loss. We optimize for 200 epochs using AdamW (learning rate 10^{-3}, weight decay 10^{-4}) and a batch size of 256. During this stage, all backbone encoders (e.g., BioCLIP, DINOv2) remain frozen. The goal is to initialize the encoder with representations that are discriminative across the full training set.

The multimodal input is passed through an MLP encoder comprising a hidden layer of size 512, batch normalization, ReLU activation, and dropout (rate 0.1). The output is projected to a 512-dimensional L2-normalized embedding. A classifier head maps this embedding to logits over C classes via a linear

layer (256 units), batch normalization, ReLU, dropout (rate 0.2), and a final projection to $\mathbb{R}^C$.

4.4.2. Stage 2: Prototypical Network Fine-Tuning

In the second stage, we adapt the encoder for few-shot classification using a prototypical network [4]. For each N-way K-shot task, class prototypes are constructed by averaging the embeddings of the K support examples per class:

$$\mu_c = \frac{1}{|S_c|} \sum_{x_i \in S_c} f(x_i)$$

Given a query embedding $f(x)$, classification is performed by computing the squared Euclidean distance to each prototype and applying a softmax over the negative distances:

$$p(y = c \mid x) = \frac{\exp\left(-\|f(x) - \mu_c\|_2^2\right)}{\sum_{c'} \exp\left(-\|f(x) - \mu_{c'}\|_2^2\right)}$$

The episodic training objective is then the average cross-entropy loss across all queries in the task:

$$\mathcal{L}_{\text{proto}} = -\frac{1}{|\mathcal{Q}|} \sum_{(\mathbf{x}, y) \in \mathcal{Q}} \log p_\phi(y \mid \mathbf{x})$$

The encoder parameters ϕ are jointly optimized using the AdamW optimizer [11] with a learning rate of 10^{-3} and weight decay of 10^{-4}. All pretrained backbones (e.g., BioCLIP, DINOv2) are kept frozen during this phase.

4.5. Episodic Training Procedure

To enable few-shot generalization, we adopt an episodic training strategy that mirrors the test-time setting. Each episode is a 5-way 5-shot task with 15 query examples per class (75 queries total), encouraging adaptation to novel classes under limited supervision. Training spans 20 meta-epochs of 200 episodes each. For every episode, the model computes class prototypes from the support set and evaluates the query set, using a classification loss based on distances to prototypes. The shared encoder and classifier are updated via AdamW with stage-1 hyperparameters, while the image and text backbones remain frozen.

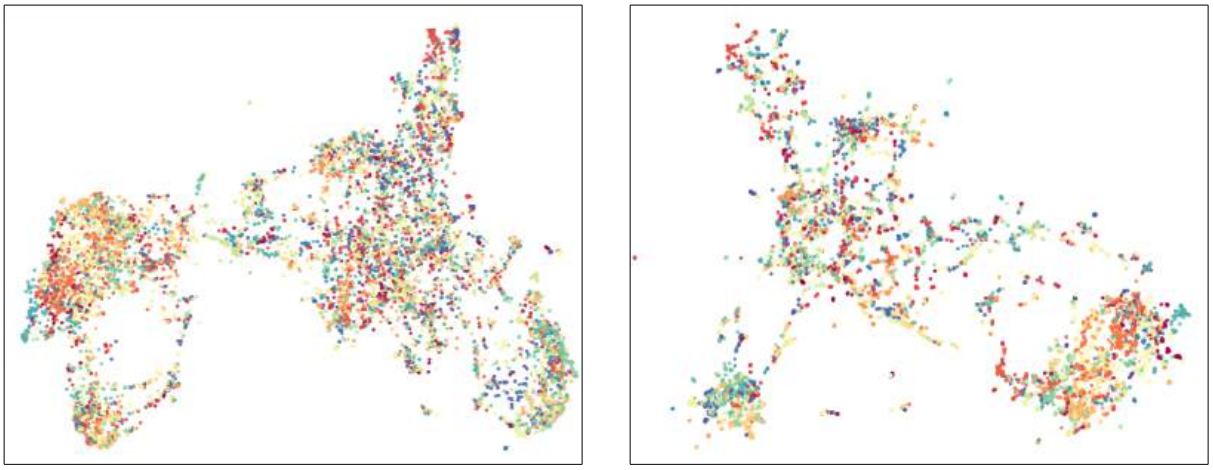
The loss is the negative log-likelihood over a softmax of distances, pushing embeddings to be discriminative and robust in data-scarce conditions. This episodic framework is inspired by prior metric-based few-shot methods such as Matching Networks [12] and Prototypical Networks [4], which have shown strong performance in low-data regimes.

4.6. Observation-Level Reranking via Prediction Aggregation

To evaluate the model at the level of fungal *observations*—sets of images of the same specimen—we aggregate image-level predictions using a weighted reranking scheme. Each image yields a top-10 list of predicted classes with confidence scores. For each observation, we collect all such predictions and apply a rank-based voting scheme with weights $w = [12, 10, 8, 7, 6, 5, 4, 3, 2, 1]$, assigning higher scores to higher-ranked classes. The aggregated scores across all images determine the observation-level top-10 prediction.

5. Results

We assess model performance using **Top**-k Accuracy (Recall@k), which measures the fraction of test samples where the true label is among the top-k predictions:



(a) Before prototypical training (b) After prototypical training

Figure 7: UMAP projections of training set embeddings before and after prototypical network fine-tuning. Prior to few-shot adaptation (left), embeddings show limited class separation. After episodic training with prototype supervision (right), embeddings become more structured and class-separable, indicating improved support-query alignment.

$$\text{Top-}k \text{ Accuracy} = \frac{1}{N} \sum_{i=1}^{N} 1(y_i \in \hat{Y}_i^k)$$

where N is the number of samples, y_i the true label, $\hat{Y}_i^k$ the top-k predictions, and $1(\cdot)$ the indicator function. We report results with $k = 5$ on both public and private test sets from FungiCLEF 2025.

5.1. Hardware

We conducted all experiments using the freely available NVIDIA Tesla P100 GPUs provided by the Kaggle platform. This hardware configuration was sufficient for extracting features from large image batches and training our few-shot models without requiring additional computational resources.

5.2. Main Evaluation

Table 3b summarizes performance across three configurations: (i) a zero-shot baseline using a single pretrained model with fused image, image-text, and metadata features, (ii) a model trained using the two-stage procedure described in Section 4, and (iii) a reranked ensemble that combines both prediction sources.

5.3. Impact of Multimodal Fusion

To isolate the effect of different input modalities, Table 3a presents Recall@5 scores using the same pretrained architecture with only a single feature modality at a time, or all three combined (image, image-text, and metadata). These results demonstrate the additive benefit of incorporating textual and contextual metadata beyond image features alone.

5.4. Reranked Ensemble Strategy

The ensemble result is obtained by merging predictions from the pretrained and trained models at the observation level using the reranking method described in Section 4.6. Specifically, predictions from both models are aggregated per observation using rank-based voting, assigning higher weights

to top-ranked predictions. This fusion strategy exploits the complementary strengths of pretrained generalization and fine-tuned specialization, yielding the best overall performance.

Table 3
Recall@5 comparison across input modalities and ensemble strategies on the FungiTastic dataset. *BioCLIP only* serves as the CL baseline provided by the challenge organizers.

(a) Modalities

Input	Public	Private
BioCLIP only	0.33185	0.26649
Image (3×)	0.30973	0.30012
+ Text	0.44690	0.44243
+ Metadata	0.46902	0.45795
All (I+T+M)	**0.46460**	**0.46183**

(b) Ensemble

Model	Public	Private
Pretrained	0.46460	0.46183
Trained	0.53539	0.50711
Ensemble	**0.57079**	**0.55498**

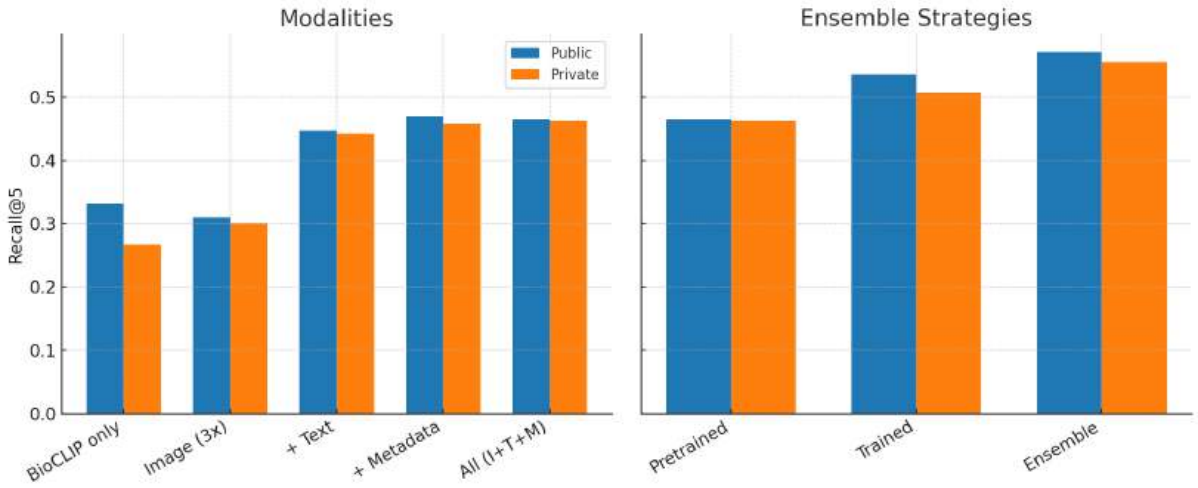


Figure 8: Recall@5 comparison for different input modalities (left) and ensemble strategies (right) on the FungiTastic dataset. Public and private split performance are shown for each configuration.

6. Ablation Study

We conducted an ablation study to assess the effectiveness of learned representations for textual and geographic information in our few-shot fungal classification pipeline.

6.1. Text Embeddings

For textual features, we replaced handcrafted rule-based and TF-IDF features with pretrained language models, including **BGE-large-en** [13], **BioCLIP**'s text encoder [7], and **e5-large** [14]. Despite their strong performance on general language understanding tasks, all models underperformed when used as sole text representations in our pipeline. These encoders failed to capture fine-grained domain-specific cues such as structured habitat descriptions or morphological terms, which are explicitly encoded by our handcrafted pipeline.

6.2. Geographic Embeddings

We also evaluated domain-adapted geographic representation models, specifically **GeoCLIP** [15] and **TaxaBind** [16], to embed geographic coordinates from the metadata. These models encode spatial and ecological priors via geolocation-aware training, but when integrated into our multimodal classification

framework, they failed to improve performance. The observed degradation is likely due to the high intra-class geographic variance and sparse sampling in the training data, which hinders the utility of learned spatial embeddings. As a result, we decided not to include geographic coordinates in the final submission.

These results underscore the importance of carefully engineered feature representations in low-resource ecological settings. While large-scale pretrained models offer generality, domain-specific heuristics currently yield more discriminative power in our few-shot fungal classification task.

7. Conclusion

We proposed a multimodal framework combining pretrained image encoders, textual descriptions, and metadata for fungal species classification on the FungiCLEF 2025 dataset. Our two-stage training with prototypical fine-tuning and observation-level reranking significantly improved few-shot performance. Multimodal fusion consistently outperformed image-only baselines, and the reranked ensemble achieved the best results. Future work will focus on enhancing metadata integration and fusion strategies to further boost accuracy in ecological image recognition.

8. Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT in order to: Improve writing style, Paraphrase and reword. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] Klarka, picekl, Fungiclef25 @ cvpr-fgvc lifeclef, https://kaggle.com/competitions/fungi-clef-2025, 2025. Kaggle.

[2] T. Berg, J. Liu, S. W. Lee, M. L. Alexander, D. W. Jacobs, P. N. Belhumeur, Birdsnap: Large-scale fine-grained visual categorization of birds, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2019–2026. doi:10.1109/CVPR.2014.259.

[3] G. Van Horn, S. Branson, R. Farrell, S. Haber, J. Barry, P. Ipeirotis, P. Perona, S. Belongie, Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 595–604.

[4] J. Snell, K. Swersky, R. S. Zemel, Prototypical networks for few-shot learning, 2017. URL: https://arxiv.org/abs/1703.05175. arXiv:1703.05175.

[5] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, 2021. URL: https://arxiv.org/abs/2103.00020. arXiv:2103.00020.

[6] X. Zhai, B. Mustafa, A. Kolesnikov, L. Beyer, Sigmoid loss for language image pre-training, 2023. URL: https://arxiv.org/abs/2303.15343. arXiv:2303.15343.

[7] S. Stevens, J. Wu, M. J. Thompson, E. G. Campolongo, C. H. Song, D. E. Carlyn, L. Dong, W. M. Dahdul, C. Stewart, T. Berger-Wolf, W.-L. Chao, Y. Su, Bioclip: A vision foundation model for the tree of life, 2024. URL: https://arxiv.org/abs/2311.18803. arXiv:2311.18803.

[8] L. Picek, K. Janouskova, V. Cermak, J. Matas, Fungitastic: A multi-modal dataset and benchmark for image categorization, 2025. URL: https://arxiv.org/abs/2408.13632. arXiv:2408.13632.

[9] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, P. Bojanowski, Dinov2: Learning robust visual features without supervision, 2024. URL: https://arxiv.org/abs/2304.07193. arXiv:2304.07193.

[10] Anthropic, Introducing claude 3.5 sonnet, 2024. URL: https://www.anthropic.com/news/claude-3-5-sonnet.

[11] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, 2019. URL: https://arxiv.org/abs/1711.05101. arXiv:1711.05101.

[12] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, D. Wierstra, Matching networks for one shot learning, 2017. URL: https://arxiv.org/abs/1606.04080. arXiv:1606.04080.

[13] S. Xiao, Z. Liu, P. Zhang, N. Muennighoff, C-pack: Packaged resources to advance general chinese embedding, 2023. arXiv:2309.07597.

[14] L. Wang, N. Yang, X. Huang, B. Jiao, L. Yang, D. Jiang, R. Majumder, F. Wei, Text embeddings by weakly-supervised contrastive pre-training, 2024. URL: https://arxiv.org/abs/2212.03533. arXiv:2212.03533.

[15] V. V. Cepeda, G. K. Nayak, M. Shah, Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization, 2023. URL: https://arxiv.org/abs/2309.16020. arXiv:2309.16020.

[16] S. Sastry, S. Khanal, A. Dhakal, A. Ahmad, N. Jacobs, Taxabind: A unified embedding space for ecological applications, 2024. URL: https://arxiv.org/abs/2411.00683. arXiv:2411.00683.