Domain Fractal Hierarchical Vision Transformer with Uncertainty Aware Active Learning for Microfossil Classification

Sangeen Khan, Tehreem Fatima, Kebede Gutema Hirpesa
Shenzhen Institutes of Advanced Technology, University of Chinese Academy of Sciences, China
Email: {sangeenkhan2662, fatima, kebede}@mails.ucas.ac.cn

***Abstract*—Microfossil images, particularly foraminifera from sediment cores, present a challenging visual recognition problem because of subtle morphological variation, overlapping shape features, strong class imbalance, and limited expert labeled data. This work proposes a Domain Fractal Hierarchical Vision Transformer (DFH ViT) combined with synthetic domain specific pretraining and uncertainty aware active learning. The core idea is to teach a Vision Transformer backbone to understand fossil like geometry using a synthetic FossilFractal dataset, then fine tune on real microfossil images from the Endless Forams family of datasets, and finally use uncertainty estimates to guide expert labeling effort where the model is most confused.**

This implementation in Google Colab realizes the full DFH ViT pipeline in PyTorch on real microfossil data. It uses the Endless Forams training set as a labeled benchmark and two downcore fossil training sets, MD022508 and MD972138, as sources of additional images and as an unlabeled pool for active learning. The notebook implements FossilFractal pretraining, hierarchical and metric fine tuning on Endless Forams, evaluation with confusion matrices and per class metrics, and an active learning loop based on Monte Carlo dropout that selects uncertain images from the MD cores. The same architecture and training code will later be extended to additional radiolarian and foraminifera collections with metadata.

***Index Terms*—Microfossils, foraminifera, Vision Transformers, active learning, metric learning, synthetic data.**

I. Introduction

Microfossils such as planktonic and benthic foraminifera play a key role in paleoceanography, paleoclimate reconstruction, and stratigraphic correlation. Automated identification and counting of these fossils can save significant expert time and improve consistency, yet it remains difficult because classes are visually similar, highly imbalanced, and often partially degraded in images.

The Endless Forams project assembled more than thirty four thousand expert annotated images of modern planktonic foraminifera across several dozen species to support taxonomic training and automated species recognition using convolutional neural networks [1]. The image based workflow was later adapted to fossil records in sediment cores, where convolutional neural networks were used to classify foraminifera from cores MD02-2508 and MD97-2138 and to perform automatic abundance analysis [2]. A derivative training dataset was released on Zenodo that provides machine learning ready cropped images for Endless Forams, MD022508, and MD972138 with text panels removed and images padded to squares [3].

Recent progress in computer vision with Vision Transformers has made it possible to learn powerful representations from images, while synthetic data and self supervised pretraining help reduce dependence on large labeled datasets. However, directly training a generic transformer on imbalanced microfossil data can still lead to poor minority class performance and unstable behavior for rare taxa. In practice, experts are also expensive and cannot label large numbers of new images in an unstructured way.

This work addresses these challenges by combining three ideas in a unified system:

- A Domain Fractal synthetic dataset that captures key fossil like patterns and can be generated at scale to support pretraining.
- A hierarchical Vision Transformer (DFH ViT) that predicts both coarse and fine labels and uses metric learning to separate visually confusable taxa.
- An uncertainty aware active learning loop that uses Monte Carlo dropout to select the most informative images for expert labeling from downcore fossil datasets.

The current Colab implementation validates the DFH ViT pipeline on the Endless Forams training set together with MD022508 and MD972138. This reduces technical risk before moving to additional microfossil collections and more complex metadata.

II. Problem Statement and Motivation

The main scientific goal is to improve class averaged accuracy and robustness in microfossil classification, especially for rare and visually similar taxa, while keeping expert labeling effort manageable.

Concrete challenges include:

- **Class imbalance and data scarcity**: Some foraminiferal species have many labeled examples, while others have very few. Direct training tends to favor majority classes and neglect rare ones.
- **Morphological similarity**: Many species share similar shell shapes, chamber arrangements, and surface textures, so a flat classifier tends to confuse these classes.
- **Limited expert time**: Experts cannot label unlimited new images. A naive labeling strategy wastes time on easy

cases while the model remains uncertain on edge cases and rare taxa.

- **Heterogeneous acquisition conditions**: Differences in imaging setup, sediment cores, and depth intervals can shift the data distribution and introduce hidden biases.

To address these challenges, the proposed system uses synthetic domain-specific pretraining to reduce data scarcity, hierarchical and metric objectives to handle class similarity, and active learning to prioritize expert effort for ambiguous fossil images in downcore datasets.

III. Objectives

The main objectives are as follows:

- Design and implement a Domain Fractal synthetic generator (FossilFractal) that produces fossil like images with controllable coarse and fine labels.
- Develop a Domain Fractal Hierarchical Vision Transformer that supports coarse and fine predictions, integrates metric learning, and optionally fuses metadata such as core identifier, depth, and age band.
- Implement and validate an uncertainty aware active learning loop that uses Monte Carlo dropout based predictive entropy to select difficult unlabeled samples from MD022508 and MD972138.
- Demonstrate the feasibility and behavior of the full pipeline in Google Colab using the Endless Forams training set as a labeled benchmark and MD022508 plus MD972138 as an unlabeled pool, with clear correspondence between the code and the proposed architecture.
- Extend the same implementation to additional microfossil datasets and evaluate performance in terms of class averaged accuracy, macro F1, confusion matrices, and active learning gain, with a focus on rare species and downcore fossil series.

IV. Related Work

Work on microfossil recognition has explored classical image processing, convolutional neural networks, and transfer learning from large scale image datasets. The Endless Forams library of modern planktonic foraminifera images has been a central benchmark for supervised classification and for training taxonomic experts [1]. Marchant et al. extended this approach to fossil records from specific sediment cores and demonstrated that automated abundance estimates can closely track manual counts over long stratigraphic intervals [2].

Vision Transformers and masked autoencoders have shown strong performance in natural image tasks and in several scientific imaging applications, while deep metric learning has been applied to foraminifera for open set recognition and visualization of morphology space [**?**]. Self supervised pretraining and domain specific synthetic data generation can reduce the reliance on large labeled datasets. Active learning and uncertainty based sampling have been widely studied in machine learning to reduce labeling cost, but they are rarely integrated in a complete microfossil recognition pipeline that includes domain specific synthetic pretraining and downcore fossil data.

The DFH ViT proposal builds on these ideas and integrates them into a single system tailored for foraminifera and similar microfossils. The current Colab implementation on Endless Forams plus MD022508 and MD972138 serves as a technical proof of concept for the training, evaluation, and active learning loop before deploying the same approach on additional cores and taxa.

V. Proposed System

A. Overview

The proposed system contains three main stages:

1) Synthetic domain specific pretraining on FossilFractal images, using multi task objectives to shape the Vision Transformer backbone toward fossil like geometry.
2) Hierarchical and metric fine tuning on real microfossil images from the Endless Forams training set, with separate heads for coarse and fine labels and a supervised contrastive loss on the embedding.
3) Uncertainty aware active learning from an unlabeled image pool built from MD022508 and MD972138, where the model queries experts only for high entropy samples.

In the Colab implementation, this pipeline is realized in PyTorch. FossilFractal provides synthetic data for pretraining. The Endless Forams training set is used as the labeled dataset for supervised fine tuning and validation. MD022508 and MD972138, which are benthic and planktonic foraminiferal training sets from individual sediment cores, are loaded as an unlabeled pool for uncertainty estimation and active sample selection. The notebook defines a single `run_training` function that handles FossilFractal pretraining, fine tuning on Endless Forams, and checkpointing of the best model. Additional cells perform evaluation and active learning experiments on the MD cores.

B. Domain Fractal Synthetic Pretraining

The FossilFractal dataset is a synthetic image generator that draws fossil like shapes with controllable structure. In the current code, simple primitives such as circles with spines, conical shells with chambers, and elongated diatom like ellipses are used. Each synthetic image receives:

- A coarse label, for example spumellarian like, nassellarian like, or diatom like.
- A fine label that encodes a synthetic cluster or subtype.

The DFH ViT model is pretrained on this dataset with a multi task loss

$$L_{\text{pre}} = \alpha L_{\text{sup}} + \beta L_{\text{MAE}} + \gamma L_{\text{aux}}, \tag{1}$$

where L_{sup} is the cross entropy loss on coarse and fine labels, L_{MAE} is a reconstruction loss on downsampled images, and L_{aux} is an auxiliary self supervised term implemented as a simple rotation like prediction derived from labels. The Colab notebook implements this pretraining loop in a function

pretrain_epoch, which is called from run_training before real data fine tuning.

As a next step, the FossilFractal generator will be extended with more realistic morphological rules and parameterized shapes that better approximate foraminiferal shells. Since the pretraining code is modular, improvements in the generator directly improve the quality of the learned backbone without changing the downstream pipeline.

C. Hierarchical DFH ViT Backbone

The core model is a Vision Transformer backbone f_θ with no classification head in the pretrained stage. After pretraining, DFH ViT adds two heads:

- A coarse head h_{coarse} that maps the embedding $z = f_\theta(x)$ to a probability distribution over coarse groups.
- A fine head h_{fine} that takes both z and the coarse logits as input and outputs probabilities over the fine classes.

The hierarchical fine tuning loss on real data is

$$L_{\text{fine}} = L_{\text{coarse}} + \lambda_1 L_{\text{fineCE}} + \lambda_2 L_{\text{metric}}, \tag{2}$$

where L_{coarse} and L_{fineCE} are cross entropy losses for the coarse and fine predictions, and L_{metric} is a supervised contrastive loss on the normalized embedding. This metric term encourages embeddings of the same class to cluster while pushing apart different classes, which is important for visually similar taxa.

In the current Colab implementation, this structure is realized as the DFHViT class. The model is first pretrained on FossilFractal and then fine tuned on the Endless Forams training set via a finetune_epoch function. The number of fine classes is inferred from the folder structure of the Endless Forams training zip, which contains one directory per species. Species are then split into two artificial coarse super classes to exercise the hierarchical heads. The code also computes validation loss, coarse accuracy, fine accuracy, and macro F1.

D. Metadata Fusion

In the fossil setting, each image can have associated metadata such as core identifier, depth, and age band. DFH ViT includes an optional metadata pathway where these values are encoded into a vector and passed through a small multilayer perceptron. The resulting metadata embedding is added to the image embedding so that self attention can use both visual and contextual information.

The current Endless Forams prototype does not yet use metadata, but the implementation already exposes a metadata_dim argument and a meta_mlp module in DFHViT. This makes the transition to microfossil data with metadata straightforward once such fields are available, since the forward pass already supports an optional metadata tensor.

E. Uncertainty Aware Active Learning

To reduce expert labeling effort and focus on the most informative samples, DFH ViT uses an uncertainty aware active learning loop. The key idea is to approximate predictive uncertainty for each unlabeled image and query experts only for high uncertainty cases.

The implementation uses Monte Carlo dropout as follows:

1) For a batch of unlabeled images from MD022508 or MD972138, run the model multiple times with dropout active and collect the predicted fine class probabilities.
2) Average the probabilities across Monte Carlo samples to obtain $\bar{p}_k(x)$ for each class k.
3) Compute the predictive entropy

$$H(x) = -\sum_k \bar{p}_k(x) \log \bar{p}_k(x). \tag{3}$$

4) Select the top K images with the highest entropy and send them for expert annotation.

In the Colab notebook, an unlabeled pool is created using EndlessForamsUnlabeledPool, which scans MD022508 and MD972138 and loads images without using their labels. Active learning utilities predictive_entropy_mc_dropout and select_uncertain_samples compute entropy scores and return the indices and entropy values of the most uncertain images. A separate evaluation cell can visualize high entropy images in a grid to allow qualitative inspection of model uncertainty. For real experiments, these selected images would be presented to experts for labeling and then added to the labeled training set for another fine tuning round.

VI. Implementation Plan

A. Datasets: Endless Forams and MD Cores

The implementation uses the Zenodo record titled Endless Foram, MD022508 and MD9712138 training datasets, which provides three machine learning ready training sets prepared for automated analysis of foraminifera fossil records [**?**], [2]. The datasets are:

- **Endless Forams training set**: This set is a derivative of the original Endless Forams image library of modern planktonic foraminifera, where the white text panel and border have been removed and images are padded to a square format suitable for convolutional and transformer models [**?**], [1]. The original Endless Forams collection contains 34,640 expert annotated images spanning 35 modern species, acquired from microscope images of sorted tests and labeled by taxonomic specialists [1]. The training zip on Zenodo groups images into species specific folders that can be used directly as a supervised classification benchmark.
- **MD022508 training set**: This set contains benthic foraminiferal images from sediment core MD02-2508 in the north eastern Pacific. It was acquired as part of the workflow used by Marchant et al. to test automated classification on downcore fossil records [2]. Images are cropped around individual tests and organized by species, and they span the last climatic cycle in the core.
- **MD972138 training set**: This set contains planktonic foraminiferal images from sediment core MD97-2138

in the western Pacific warm pool region [2]. As with MD022508, images are cropped and labeled by species and cover a long stratigraphic interval, making them suitable for testing automated abundance reconstruction in fossil assemblages.

These training sets were originally prepared to support convolutional neural network based classification and abundance estimation. The DFH ViT implementation reuses them as follows: Endless Forams is treated as the main labeled dataset for supervised fine tuning and validation, while MD022508 and MD972138 are loaded as unlabeled pools for uncertainty estimation and active sample selection. Because the Zenodo packages are organized by species folders and already preprocessed, they can be used in Colab with minimal additional data cleaning.

B. Current Prototype in Colab

The current implementation in Google Colab covers the full DFH ViT pipeline and is aligned with this proposal:

- A `SyntheticFossilFractalDataset` class that generates synthetic FossilFractal images with simple hierarchical labels.
- A `DFHViT` model that includes a Vision Transformer backbone, hierarchical heads for coarse and fine classification, a reconstruction head, and an auxiliary head, with an optional metadata branch.
- Pretraining code that applies the multi task loss L_{pre} on FossilFractal before real data fine tuning.
- Fine tuning code that trains DFH ViT on the Endless Forams training set with hierarchical and metric losses and stores the best checkpoint according to macro F1 on the validation split.
- Evaluation utilities that compute overall metrics and a unified evaluation cell that draws confusion matrices, per class accuracy bar plots, and example prediction grids from the validation set.
- Active learning utilities that compute Monte Carlo dropout based entropy, select top K uncertain images from the MD022508 and MD972138 pools, and simulate further fine tuning steps after adding such samples to the labeled data.

The Colab notebook demonstrates that the DFH ViT architecture, training objectives, evaluation, and active learning loop run end to end on real foraminifera datasets that are widely used in the literature. The code is organized in modular cells so that additional cores or radiolarian datasets can be incorporated by adding new dataset wrappers without changing the core model or learning logic.

C. Transition to Additional Microfossil Data

The next stage is to adapt the prototype to further microfossil datasets beyond the Endless Forams and MD cores. Planned steps include:

- Implement a `MicrofossilDataset` class that reads generic metadata CSV files with columns such as image path, coarse label, fine label, core identifier, depth, and age band, and reuse the same transforms and DataLoader setup.
- Adjust configuration to use the true number of fine classes and coarse groups for each new dataset, while reusing the same DFH ViT definition, pretraining weights, and loss formulation.
- Enable metadata fusion by passing metadata tensors into the DFH ViT model, tuning the metadata embedding, and evaluating the impact of context on classification accuracy.
- Run FossilFractal pretraining and Endless Forams fine tuning as a base stage, then continue fine tuning on new datasets to study transferability and domain shift.
- Create unlabeled pools from additional cores and apply the same active learning selection logic used for MD022508 and MD972138 to guide expert labeling.

VII. Evaluation Plan

Evaluation will focus on both performance and labeling efficiency.

A. Performance Metrics

The main metrics are:

- Fine class accuracy and macro F1 score over all species.
- Coarse group accuracy.
- Per class accuracy, especially for rare taxa in the Endless Forams and MD training sets.
- Confusion matrices to visualize which species are commonly confused.

Baselines will include:

- A standard Vision Transformer or convolutional network fine tuned directly on the Endless Forams training set without synthetic pretraining.
- DFH ViT without FossilFractal pretraining.
- DFH ViT without the metric loss term.
- DFH ViT without active learning, using a fixed random labeled subset of equal size when integrating additional samples from MD022508 and MD972138.

B. Active Learning Evaluation

For active learning, the key questions are:

- How classification performance improves as a function of the number of labeled images when using uncertainty based selection from MD022508 and MD972138 compared with random selection.
- How many labeled samples per species are needed to reach a desired accuracy threshold on the Endless Forams validation split and on downcore fossil subsets.
- Whether high entropy selections correspond to visually ambiguous or under represented taxa as perceived by experts inspecting the selected MD core images.

Experiments will simulate active learning by starting from a small labeled seed set, adding batches of high entropy images from the MD pools, retraining or fine tuning the model, and measuring performance after each round. This procedure

is already implemented and tested in the Colab notebook and will be repeated under different sampling strategies and hyperparameters.

VIII. Expected Contributions

The expected contributions of this work are:

- A Domain Fractal synthetic generator tailored to fossil like morphology that can support data efficient pretraining of Vision Transformers for microfossil classification.
- A hierarchical Vision Transformer architecture (DFH ViT) that jointly models coarse and fine foraminiferal labels and uses metric learning to reduce confusion between visually similar taxa.
- An integrated uncertainty aware active learning loop that directs expert attention to the most informative microfossil images from downcore fossil datasets and can be implemented in practical labeling workflows.
- A complete, open implementation in Colab that demonstrates the approach on the Endless Forams training set plus the MD022508 and MD972138 cores, with a clear path to apply it to additional microfossil datasets with metadata.

IX. Conclusion

This paper presents an updated and implementable plan for microfossil classification using a Domain Fractal Hierarchical Vision Transformer and uncertainty aware active learning. The current Colab prototype shows that the architecture, training objectives, evaluation, and active learning loop can be implemented and tested on real foraminifera image datasets from the Endless Forams initiative and MD cores. Future work will refine the synthetic generator, expand to additional datasets, integrate metadata, and evaluate performance and labeling efficiency in collaboration with domain experts.

References

[1] A. Y. Hsiang, A. Brombacher, M. C. Rillo, M. J. Mleneck-Vautravers, S. Conn, S. Lordsmith, A. Jentzen *et al.*, "Endless forams: ¿34,000 modern planktonic foraminiferal images for taxonomic training and automated species recognition using convolutional neural networks," *Paleoceanography and Paleoclimatology*, vol. 34, no. 7, pp. 1157–1177, 2019.

[2] R. Marchant, M. Tetard, A. Pratiwi, M. Adebayo, and T. de Garidel-Thoron, "Automated analysis of foraminifera fossil records by image classification using a convolutional neural network," *Journal of Micropalaeontology*, vol. 39, no. 2, pp. 183–202, 2020.

[3] ——, "Endless foram, md022508 and md9712138 training datasets," 2020, machine-learning-ready training sets for Endless Forams, MD02-2508, and MD97-2138. [Online]. Available: https://doi.org/10.5281/zenodo.3996436