G-025

Classification of Fungal Images Using Hybrid ViT-GNN

閻 佳玉‡　　鎌田 清一郎‡
JIAYU YAN　Sei-ichiro Kamata

1. Abstract

Microscopic analysis remains the clinical gold standard for fungal infection diagnosis, yet it is time-consuming and prone to subjectivity. To address this, we propose a hybrid deep learning framework that integrates a structure-guided Vision Transformer (StruViT) with a topology-aware Graph Neural Network (KNNSGAT) for robust fungal image classification. StruViT enhances visual feature extraction by incorporating structural priors from cell graphs into its self-attention mechanism, while KNNSGAT leverages GraphSAGE and GAT to model inter-cell relationships from mask-derived graphs. Experimental results on the high-resolution DIFaS dataset show that our approach outperforms conventional CNN and ViT models, especially in distinguishing morphologically similar species. This study demonstrates the effectiveness of combining global semantic encoding with fine-grained structural modeling for accurate and automated fungal diagnosis.

2. Introduction

Fungal infections remain a major public health concern, affecting over one billion individuals worldwide and resulting in approximately 1.5 million fatalities annually [1][2]. Microscopic examination, although considered the clinical gold standard, is labor-intensive, time-consuming, and highly dependent on the examiner's expertise. These challenges have spurred increasing interest in developing automated fungal diagnostic systems based on deep learning, aiming to improve diagnostic efficiency and objectivity.

Deep learning methods have been increasingly applied to fungal image classification. Convolutional Neural Networks (CNNs) have shown promise [3–5], but their limited receptive fields make it difficult to distinguish morphologically similar species. Vision Transformers (ViT) [6] leverage global self-attention mechanisms and offer improved semantic modeling, yet they lack explicit mechanisms for capturing cell-level structural dependencies [7]. Meanwhile, Graph Neural Networks (GNNs) [8] are effective at encoding topological relationships among segmented regions, but they often fall short in modeling global visual semantics.

To address these challenges, we propose a hybrid architecture that incorporates graph-guided structural priors into the visual representation through cross-attention, enabling more accurate semantic modeling:

(1) We inject graph-derived structural bias into the self-attention layers of a Transformer encoder to enhance its sensitivity to spatial relationships between fungal cells, thereby improving semantic representation at the patch level.

(2) We construct cell-level graphs from segmentation masks using K-nearest neighbor sampling and apply graph neural aggregation with attention to capture topological dependencies among cell regions.

(3) We design a fusion mechanism based on cross-attention that integrates global visual tokens and local structural features, enabling the model to generate a unified representation for final classification.

3. Related Work

Existing studies on fungal image classification can be broadly categorized into the following four types: traditional CNN-based supervised methods, transformer-based global modeling approaches, graph-based structural modeling techniques, and meta-learning and unsupervised learning frameworks.

Several early approaches employed standard convolutional neural networks (CNNs) for fungal classification. Zieliński et al. [3] used AlexNet combined with Fisher Vector encoding and an SVM classifier. Sopo et al. [4] and Yilmaz et al. [5] adopted VGG16 and InceptionV3, respectively, demonstrating basic classification feasibility on the DIFaS dataset. In addition, Cagatan et al. [9] proposed a VGG16-based deep learning model specifically for detecting *C. neoformans*, showing potential for automated fungal diagnosis. While effective at capturing local patterns, these models often struggle with morphological ambiguity due to their limited receptive fields.

Recent studies have explored the use of transformer architectures for fungal image analysis. Ahmed et al. [6] combined vision transformers with CNNs to leverage global semantic representations. Vision Transformers (ViT) [7] have shown superiority in capturing long-range dependencies; however, they lack inductive bias toward spatial structure, which is crucial in biological microscopy.

To handle data scarcity, Rawat et al. [10] introduced a meta-learning strategy using MeFunX for few-shot fungal classification. In parallel, Liu et al. [11] proposed an unsupervised classification pipeline utilizing ConvNeXt feature extraction, UMAP-based dimensionality reduction, and clustering ensemble voting. This method achieved a classification accuracy of 94.1% on the DIFaS dataset, outperforming several supervised baselines. Nevertheless, the post-labeling process limits its real-time applicability and interpretability.

Graph-based approaches attempt to model inter-cell spatial relationships. Pooya Sajjadi et al. [12] proposed DPGN, an optimized graph neural network based on ResNet-50 backbones, achieving improved accuracy by encoding local topological dependencies. However, such models alone may miss holistic image semantics when used without global contextual information.

In contrast to the above methods, our work proposes a unified deep learning framework that integrates both global visual context and structural cell-level information. Specifically, we design a structure-aware transformer backbone and a topology-guided GNN encoder, which are fused via a cross-attention mechanism to enable robust fungal image classification in an end-to-end manner.

4. Proposed Method

4.1 StruViT

We adopt a ViT-like architecture as the visual backbone. As illustrated in Figure 1, the input microscopy image is resized and partitioned into non-overlapping patches, which are embedded and passed through a stack of Transformer encoder layers. To enhance structural awareness, we inject an additional structure token into the input sequence.

This token is derived from the corresponding segmentation mask, which highlights fungal cells. From the mask, we extract region-level features such as the number of cells, average size, compactness, and spatial dispersion. These features are concatenated and projected through a linear layer to form a dense vector with the same dimension as a patch embedding.

The structure token is inserted alongside the [CLS] token and patch embeddings at the input layer. It participates in self-attention, allowing the Transformer to encode global semantic features while being guided by biological structural cues. The final output of the [CLS] and structure tokens is used as the the final visual representation for downstream classification.

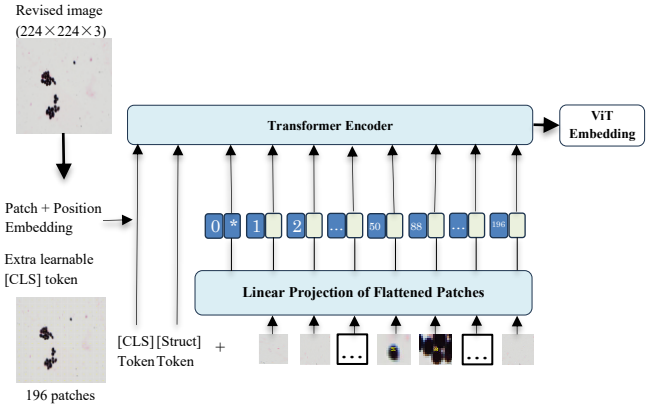

Figure 1. Architecture of the structure-guided Vision Transformer (StruViT). The input image is resized and divided into 196 non-overlapping patches. A [CLS] token and a structure-guided [Struct] token, derived from the corresponding segmentation mask, are appended to the patch sequence. All tokens are embedded and passed through the Transformer encoder, which yields a global semantic representation for classification.

$s = MLP([\mu_{area}, \mu_{perimeter}, compactness, dispersion], \dots) \in R^D$

where s denotes the structure token, and D is the dimension of the patch embedding used in the Vision Transformer.

4.2 KNNSGAT

To effectively model the spatial organization of fungal cells, we construct a cell-level graph from the segmentation mask corresponding to each microscopy image. In this graph, each connected region identified as a fungal cell is treated as a node, and its centroid is calculated based on the spatial distribution of pixels. Edges are established using a K-nearest neighbor (KNN) algorithm, where each node is connected to its closest neighbors in Euclidean space. This strategy enables the graph to reflect local spatial relationships while maintaining a sparse and computationally tractable structure.

As shown in Figure 2, for each node, we extract a set of morphological features from the corresponding cell region. These include the area and perimeter of the segmented region, compactness (defined as perimeter squared divided by area), mean and standard deviation of pixel intensities within the region, and normalized coordinates of the centroid. These features are concatenated to form a node attribute vector, which encapsulates both geometric and photometric properties relevant to fungal morphology.

To encode the graph-structured data, we design a dual-path graph neural network composed two complementary aggregation mechanisms. One employs GraphSAGE, which captures local neighborhood information through learnable feature aggregation. The other path utilizes a Graph Attention Network (GAT), allowing the model to assign adaptive weights to neighboring nodes based on their feature relevance. The outputs of these two branches are concatenated and subsequently pooled across the entire graph using a combination of mean and max pooling operations. This process yields a graph-level embedding that summarizes the structural characteristics of the fungal cells within the image. The resulting representation is then passed to the fusion module for integration with visual features extracted by the Vision Transformer.

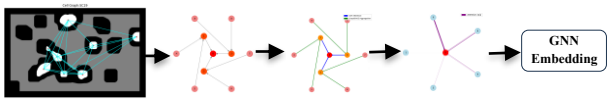

Figure 2. Pipeline of cell graph construction and GNN-based structural embedding. Each connected region in the segmentation mask is treated as a node, and edges are formed based on spatial proximity using a k-nearest neighbor (KNN) algorithm. Node features are extracted from the morphological and intensity characteristics of each cell. The resulting graph is processed by a dual-path GNN to obtain a global structure-aware embedding for classification.

4.3 Cross-Attention Fusion

While the Vision Transformer and the Graph Neural Network extract complementary features from the same microscopy image—namely, global semantics and local structural representations—their effective integration is critical for achieving robust classification. To this end, we propose a cross-attention-based fusion mechanism that enables the model to effectively align and integrate the two modalities.

In the proposed design, the graph-level embedding generated by the dual-path GNN encoder is treated as the query, while the output tokens from the Vision Transformer, specifically the [CLS] token and the structure token, are concatenated and used as the key and value. This asymmetric cross-attention formulation allows the

structural representation to selectively attend to the global visual context, enabling it to modulate its interpretation based on high-level semantic cues.

The attention output is subsequently passed through a multi-layer perceptron (MLP) consisting of fully connected layers with non-linear activation, followed by a softmax classifier. This fused representation captures both visual semantics and spatial structure, enabling robust fungal classification.

$$z = softmax\left(\frac{qK^T}{\sqrt{d_k}}\right)V$$

where $q \in R^{1\times d}$ denotes the graph-level structural embedding obtained from the GNN encoder, and $K, V \in R^{2\times d}$ are the key and value matrices constructed by concatenating the [CLS] token and the [Struct] token output from the Vision Transformer. z is the fused representation passed to the MLP classifier.

4.4 Overall Architecture

The proposed framework consists of two parallel processing branches designed to extract complementary information from fungal microscopy images. The first branch employs a structure-guided Vision Transformer, in which the input image is divided into patches and embedded into a sequence that includes a structure token derived from the corresponding segmentation mask. This token encodes cell-level statistical features and enables the Transformer to incorporate structural context during self-attention computation. The overall ViT-GNN fusion workflow is illustrated in Figure 3.

The second branch focuses on modeling topological relationships among fungal cells. It constructs a cell-level graph from the segmentation mask, where each node represents a segmented cell region with morphological attributes, and edges are formed using a K-nearest neighbor algorithm. A dual-path graph neural network, combining GraphSAGE and GAT modules, is used to extract a comprehensive graph-level representation.

These two modalities are subsequently integrated via a cross-attention fusion module, which allows the structural embedding to dynamically attend to visual features. The resulting fused representation captures both global appearance and local structure, and is passed to a classifier for final prediction. This architecture enables robust fungal classification by effectively combining semantic and structural cues.

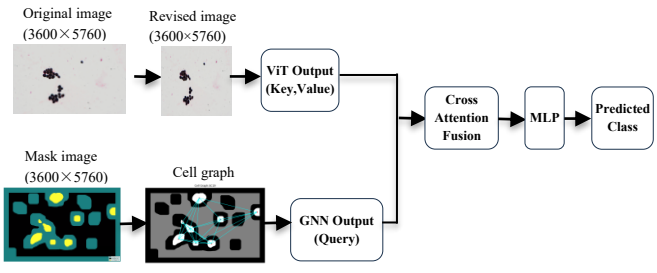

Figure 3. Overall workflow of the proposed ViT-GNN hybrid classification model. The original microscopy image is processed into a revised image for global feature extraction via a Vision Transformer (ViT), producing a [CLS] token output used as key and value. Simultaneously, the corresponding mask is used to construct a cell-level graph, which is encoded by a Graph Neural Network (GNN) to produce a structural query embedding. These two modalities are integrated through a cross-attention fusion module and classified via an MLP.

5. Experiments

5.1 Datasets

We evaluate our method on the Digital Images of Fungal Species (DIFaS) dataset [3], a publicly available dataset for fungal microscopy image classification. DIFaS comprises 180 grayscale microscopy images with a resolution of (3600×5760) pixels and 16-bit depth, covering nine clinically relevant fungal genera: CA, CG, CL, CN, CP, CT, MF, SB, and SC.

Each species includes two Gram-stained slide preparations, and multiple images are acquired per preparation, typically around 20 images per class. All images are accompanied by pixel-level segmentation masks that delineate fungal structures. The dataset is commonly used for both supervised and unsupervised classification tasks. Representative examples of each fungal species are shown in Figure 4.

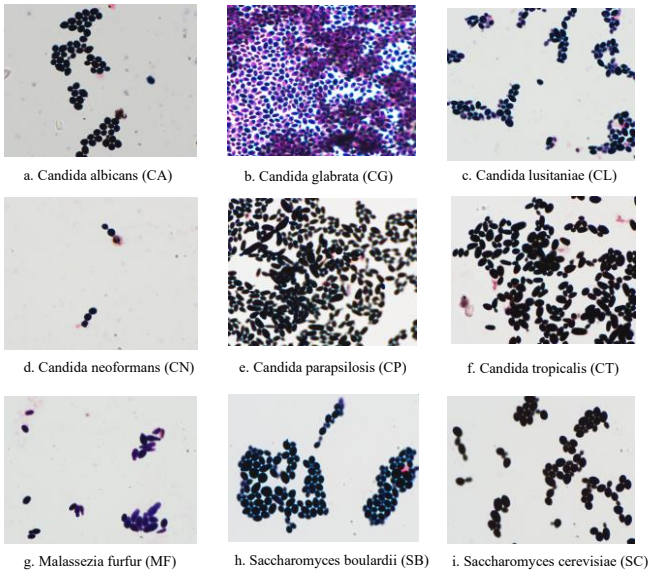

Figure 4. Sample images from the DIFaS database.

5.2 Implementation Details

All experiments were conducted using Python 3.8.20 and PyTorch 2.0.1 with CUDA 11.7. The models were trained on a workstation equipped with an Intel Core i7-9700K CPU and an NVIDIA RTX 3080 GPU.

We trained all models using the Adam optimizer (learning rate = 3×10^{-5}) and the cross-entropy loss. A batch size of 8 was used, and a fixed seed (SEED = 108) ensured reproducibility across data splitting and training. The dataset was split into training and test sets with an 80:20 ratio, while maintaining a minimum of 100 test samples per class to ensure class balance and robust evaluation.

5.3 Comparisons with State-of-the-art Methods

We compare the performance of our proposed hybrid model with several existing fungal image classification methods reported

in prior literature. As shown in Table 1, classical handcrafted pipelines such as ResNet18 combined with Bag-of-Words (BoW) features and Random Forest (RF) or SVM classifiers achieve competitive baseline performance, with accuracy values of 87.2% and 85.6%, respectively [5].

Our proposed method achieves an accuracy of 88.37%, outperforming these conventional approaches. The improvement can be attributed to the integration of global semantic representations from the Vision Transformer and local topological features from the graph neural network. This result demonstrates the effectiveness of combining modality-specific representations for fine-grained fungal classification.

Table 1. Classification accuracy of existing methods and the proposed model on the DIFaS dataset.

Method	Accuracy
ResNet18 + BoW + RF[5]	87.2±1.7%
ResNet18 + BoW + SVM[5]	85.60±2.2%
Our Proposed Method	88.37

5.4 Ablation Study

To assess the individual contribution of the visual and structural components in our model, we conduct ablation experiments by selectively disabling each branch. As shown in Table 2, the “Base + ViT only” configuration, which retains the visual backbone but removes the GNN, achieves an accuracy of 82.59%. Conversely, the “Base + GNN only” configuration, which excludes the Vision Transformer and uses only graph-based features, yields a lower accuracy of 75.31%.

When both branches are combined (“Base + ViT-GNN”), the model achieves the highest accuracy of 88.30%, demonstrating a clear performance gain from integrating global visual features and local structural information. These results confirm that the two modalities are complementary and that their combination is essential for accurate fungal classification.

Table 2. Ablation study evaluating the contribution of ViT and GNN components.

Method	Accuracy
Base + ViT only	82.59%
Base + GNN only	75.31%
Base + ViT-GNN	88.37%

6. Conclusion

In this study, we proposed a hybrid deep learning framework that integrates a structure-guided Vision Transformer with a topology-aware graph neural network for fungal image classification. The visual branch captures global semantic information from microscopy images, while the graph branch encodes local structural relationships among fungal cells. A cross-attention fusion module enables effective interaction between the two modalities.

Experimental results on the DIFaS dataset demonstrate that our method outperforms both ViT-only and GNN-only baselines, achieving improved classification accuracy. The ablation study further confirms the complementary nature of visual and structural features.

In future work, we plan to extend this framework to multi-class semi-supervised learning and explore its applicability to other biomedical imaging domains.

Acknowledgement

The authors declare no conflict of interest. Jiayu Yan: Conceptualization, Methodology, Research, Experiment, Validation, Writing—Original Draft. Sei-ichiro Kamata: Conceptualization, Supervision, Review & Editing. I would like to express my sincere gratitude to my supervisor Sei-ichiro Kamata for his continuous guidance and valuable feedback throughout this research.

References

[1] Denning, David W. Global incidence and mortality of severe fungal disease, *The Lancet Infectious Diseases*, Volume 24, Issue 7, e428–e438. https://doi.org/10.1016/S1473-3099(23)00692-8
[2] Howell SA. Dermatopathology and the Diagnosis of Fungal Infections. *Br J Biomed Sci*. 2023 Jun 7;80:11314. https://doi.org/10.3389/bjbs.2023.11314.
[3] B. Zieliski, A. Sroka-Oleksiak, D. Rymarczyk, A. Piekarczyk, M. Brzychczy-Woch. Deep learning approach to describe and classify fungi microscopic images, PLoS ONE, 15 (2020). https://doi.org/10.1371/journal.pone.0234806.
[4] C.J.P. Sopo, F. Hajati, S. Gheisari. DeFungi: direct mycological examination of microscopic fungi images, (2021). http://arxiv.org/abs/2109.07322.
[5] A. Yilmaz, F. Goktay, R. Varol, G. Gencoglan, H. Uvet. Deep convolutional neural networks for onychomycosis detection. (2021). http://arxiv.org/abs/2106.16139.
[6] Ahmed, Sheikh & Haque, A. H. M. Osama. (2023). Microscopic Fungi Classification Using Vision Transformer Guided by Transfer Learning Approach. https://doi.org/10.1109/ICCIT60459.2023.10441604
[7] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *International Conference on Learning Representations (ICLR)*, 2021.
[8] Hamilton W, Ying Z, Leskovec J. Inductive representation learning on large graphs. *Advances in Neural Information Processing Systems*, 2017, 30.
[9] A. Seyer Cagatan, M. Taiwo Mustapha, C. Bagkur, T. Sanlidag, D.U. Ozsahin. An alternative diagnostic method for C. neoformans: preliminary results of deep learning based detection model. *Diagnostics*, 13 (2023). https://doi.org/10.3390/diagnostics13010081.
[10] Shubhankar Rawat, Bhanvi Bisht, Virender Bisht, Nitin Rawat, Aditya Rawat. MeFunX: A novel meta-learning-based deep learning architecture to detect fungal infection directly from microscopic images. *Franklin Open*, Volume 6, 2024, 100069. https://doi.org/10.1016/j.fraope.2023.100069.
[11] Liu Z, Zhang F, Cheng L, Deng H, Yang X, Zhang Z, Zhou C. Simple but Effective Unsupervised Classification for Specified Domain Images: A Case Study on Fungi Images. *arXiv preprint* arXiv:2311.08995, 2023.
[12] Pooya Sajjadi, Farshid Hajati, Alireza Rezaee, Shahadat Uddin. Classification of direct microscopic fungi images using optimized graph networks. *Biomedical Signal Processing and Control*, Volume 109, 2025, 108035. https://doi.org/10.1016/j.bspc.2025.108035.