vHeat-HMCD: Hierarchical Multi-scale Contrastive Discrimination for Electron Microscopy Pollen Image Classification

1st Hongyu Song
College of
Intelligent Science
and Technology
Inner Mongolia
University of Technology
Hohhot, China
20241800119@imut.edu.cn

2nd Shi Bao*
School of
Information Engineering
Inner Mongolia
University of Technology
Hohhot, China
kshibao@imut.edu.cn

3rd Yatu Ji
College of
Intelligent Science
and Technology
Inner Mongolia
University of Technology
Hohhot, China
mljyt@imut.edu.cn

4th Min Lu
College of
Intelligent Science
and Technology
Inner Mongolia
University of Technology
Hohhot, China
cslumin@imut.edu.cn

***Abstract*—Pollen grain identification is pivotal in diverse disciplines ranging from palynology to forensic science, yet traditional manual examination remains time-consuming and labor-intensive. Although automated recognition has emerged as a solution, it remains challenged by subtle inter-class morphological similarities and significant background interference in electron microscopy images. While recent physics-inspired backbones like vHeat offer efficient global modeling, they often struggle to distinguish discriminative foreground details from complex background noise. To address this, we propose integrating a Hierarchical Multi-scale Contrastive Discrimination (HMCD) framework into the vHeat backbone to synergize efficient global modeling with robust noise filtering. Specifically, we design a Hierarchical Multi-scale Background Suppression (HMBS) module equipped with an Uncertainty-Aware Adaptive Thresholding (UAAT) mechanism, which dynamically adjusts suppression intensity based on predictive uncertainty to preserve fine-grained edge details. Furthermore, we introduce a Contrastive Feature Decoupling (CFD) module to enhance feature discriminability. Within CFD, a Dynamic Foreground Discovery (DFD) mechanism adaptively localizes foreground regions via learnable differentiable thresholding, while a Cross-Scale Contrastive Consistency (CSCC) strategy enforces semantic alignment across hierarchical stages. Extensive experiments on a large-scale self-constructed dataset and three public benchmarks demonstrate that our proposed framework significantly outperforms state-of-the-art methods, achieving superior accuracy and robustness in electron microscopy pollen image classification.**

***Index Terms*—Pollen image classification, electron microscopy, vHeat-HMCD, background suppression, contrastive learning**

I. Introduction

Pollen grain identification is pivotal in diverse scientific disciplines, ranging from palynology and climate research to allergy forecasting and forensic science. [1]–[3] Traditional identification methods rely heavily on manual examination by trained experts via optical or electron microscopy—a process that is not only time-consuming and labor-intensive but also prone to inter-observer variability. With the rapid advent of deep learning, automated pollen recognition has emerged as a robust solution to circumvent these bottlenecks.

*Corresponding author: kshibao@imut.edu.cn

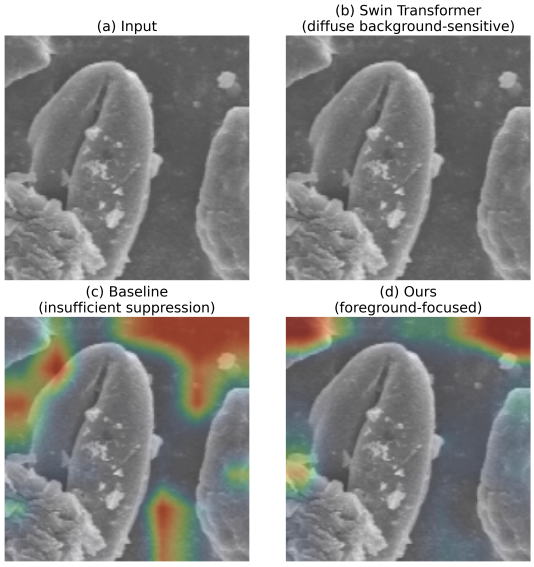

Fig. 1. **Visualization of Class Activation Maps (CAMs) comparing different approaches.** (a) Original electron microscopy image of a pollen grain. (b) Activation map from a standard Vision Transformer, exhibiting diffuse responses sensitive to background textures. (c) Activation map from the vHeat backbone, showing improved but still imprecise localization with residual background noise. (d) Activation map from our **HMCD** framework. By integrating Hierarchical Multi-scale Background Suppression (HMBS) and Contrastive Feature Decoupling (CFD), our method effectively suppresses background interference and precisely focuses on discriminative morphological details.

Despite this progress, pollen image classification poses unique challenges distinct from generic visual tasks. Different pollen species often exhibit highly similar morphological traits, resulting in subtle inter-class differences; meanwhile, a single pollen type may demonstrate significant intra-class variations driven by developmental stages, spatial orientations, and imaging conditions. Early methods predominantly relied on Convolutional Neural Networks (CNNs) [4]–[6], focusing on extracting local discriminative features to capture subtle morphological differences. With the emergence of Vision

Transformers (ViTs) [7], [8], researchers began leveraging self-attention for global context; however, ViTs often suffer from high computational complexity. To balance efficiency and performance, advanced architectures like Mamba [9], [10] and the physics-inspired vHeat backbone [11] have recently been introduced, offering new avenues for efficient global modeling.

However, applying these methods to microscopic pollen images reveals notable limitations. As illustrated in Fig. 1, standard Transformer-based models tend to produce diffuse or weakly localized responses with high sensitivity to background textures (see Fig. 1b). Similarly, vHeat also exhibits dependency on background features (see Fig. 1c). Such background interference obscures truly discriminative pollen structures, thereby hindering fine-grained recognition. To address this, we propose a Hierarchical Multi-scale Contrastive Discrimination (HMCD) framework specifically designed for fine-grained pollen recognition under electron microscopy. We integrate HMCD into the vHeat backbone to leverage its efficient global modeling capabilities while addressing its sensitivity to background noise. HMCD comprises two tightly coupled modules: Hierarchical Multi-scale Background Suppression (HMBS), which suppresses interference across feature stages, and Contrastive Feature Decoupling (CFD), which enhances discriminative representation learning through adaptive foreground localization. As shown in Fig. 1d, these components enable our framework to effectively mitigate background noise while accurately capturing scale-invariant morphological characteristics. Our main contributions are summarized as follows:

- We propose a Hierarchical Multi-scale Background Suppression (HMBS) module equipped with an uncertainty-aware adaptive thresholding mechanism to dynamically filter background interference based on predictive confidence.
- We develop a Contrastive Feature Decoupling (CFD) module that integrates dynamic foreground discovery with cross-scale contrastive consistency to enforce robust semantic alignment across different feature scales.
- Extensive experiments conducted on our self-constructed dataset (Pollen149) and three public benchmarks demonstrate that the proposed framework achieves competitive performance.

II. Related Work

A. Evolution of Visual Backbones

The evolution of visual backbones constitutes the foundation of modern computer vision. LeNet [12] pioneered this era, while AlexNet [4] subsequently established Convolutional Neural Networks (CNNs) as the mainstream solution for visual tasks. Following this breakthrough, numerous studies focused on developing deeper and more effective network architectures to push the boundaries of performance [5], [6], [13]–[18]. Concurrently, another stream of research dedicated itself to improving the convolution layers to enhance computational efficiency and feature extraction [19]–[25].

Inspired by the substantial success of Transformers [26] in NLP, ViT [7] introduced a pure self-attention mechanism to computer vision. To address the limitations of the original ViT, one stream of research introduced hierarchical pyramid architectures, enabling the model to extract multi-scale features critical for dense predictions [8], [27], [28]. Another stream focuses on optimizing the attention mechanism itself, employing strategies such as window-based or sparse attention to reduce the quadratic computational complexity [29]–[33].

Recently, Mamba architectures based on State Space Models (SSMs) have emerged as a promising alternative [34]. By incorporating mechanisms such as selective scanning and bidirectional processing, these models maintain linear computational complexity while effectively capturing long-range dependencies in non-causal visual data [9], [10], [35], [36]. Most recently, the physics-inspired vHeat was introduced, modeling information propagation as thermal energy diffusion to achieve global receptive fields with efficiency superior to self-attention [11].

B. Pollen Grain Recognition

Automated pollen recognition represents a critical domain in biological image analysis, with broad applications in palynology, climate research, allergy forecasting, and forensic science. Early methods primarily relied on manual feature engineering, utilizing texture descriptors such as Local Binary Patterns (LBP) and Gray-Level Co-occurrence Matrix (GLCM) combined with classifiers like SVMs [37]. However, these methods depend heavily on manual parameter tuning and struggle to generalize across diverse species and imaging conditions.

In recent years, with the rapid development of deep learning, it has become the mainstream approach for automated pollen image classification. Most approaches employ transfer learning [38], fine-tuning CNN backbones (e.g., ResNet, DenseNet) pre-trained on ImageNet to extract high-level semantic features from specialized datasets such as POLLEN73S [39] and Pollen23E [40]. Several recent studies have introduced attention mechanisms: RCANet [41] embeds a cognitive attention module (VDAB) inspired by the human visual dual-stream hypothesis into ResNet18; APFA-Net [42] proposes an attention-guided feature aggregation framework; and HieraEdgeNet [43] introduces multi-scale fusion of frequency-channel and spatial-channel attention to address edge blur and complex backgrounds. Additionally, hybrid architectures such as RESwinT [44] combine ResNet with Swin Transformer's parallel attention streams to capture both local textures and global dependencies. Although some methods have begun to address background interference, systematic solutions for the high background noise inherent in electron microscopy images remain lacking.

C. Contrastive Representation Learning

Contrastive learning has emerged as a dominant paradigm for self-supervised representation learning. Seminal works such as SimCLR established the importance of strong data augmentation and non-linear projection heads for instance discrimination [45], [46]. To address the memory constraints

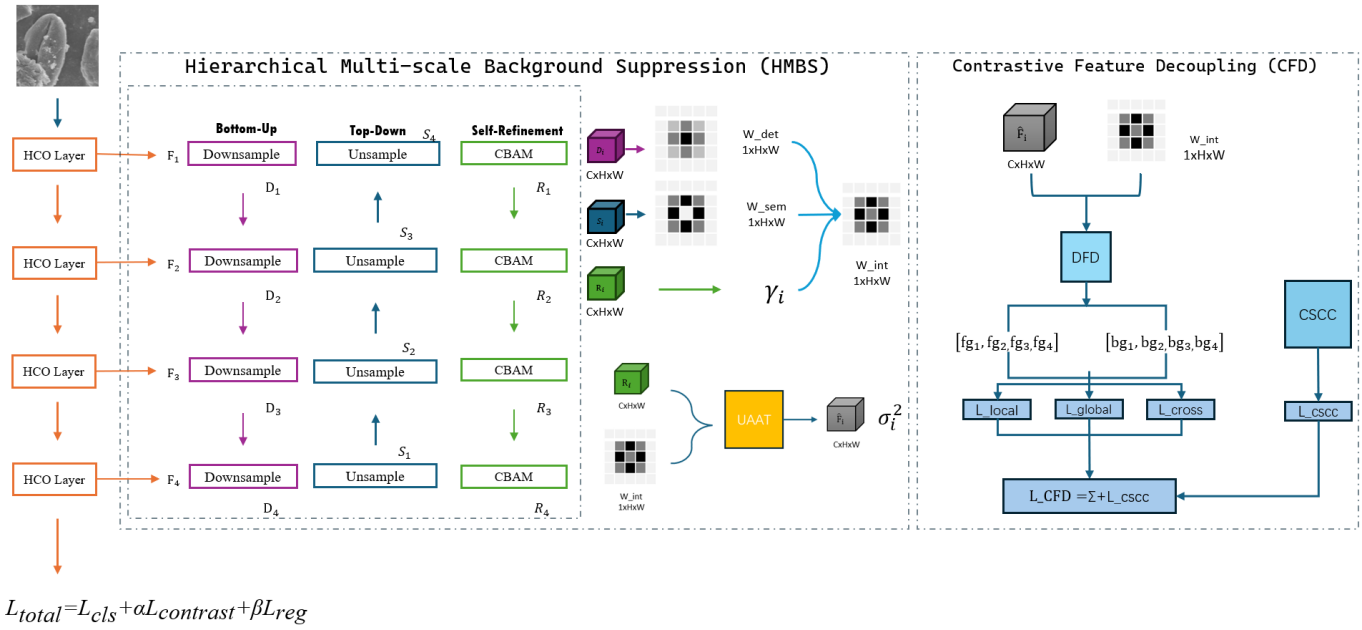

Fig. 2. Overall architecture of vHeat-HMCD. An input image is fed into the vHeat backbone to extract multi-scale features from four stages. These features are then passed to HMBS (Hierarchical Multi-scale Background Suppression) to suppress background noise and enhance foreground representations. The enhanced features are subsequently processed by CFD (Contrastive Feature Decoupling) for discriminative representation learning, and finally sent to the classification head to produce the final prediction.

of large batch sizes, MoCo introduced a momentum encoder and a dynamic dictionary look-up mechanism [47], [48].

Following these foundational frameworks, one stream of research focused on eliminating the reliance on negative pairs. Architectures like BYOL [49] and SimSiam [50] utilized asymmetric networks and stop-gradient operations to prevent mode collapse without explicit negative samples. Another stream explores prototype-based learning, where methods like SwAV enforce consistency between cluster assignments rather than individual features [51].

Recently, with the rise of Transformers, approaches such as MoCoV3 [52] and DINO [53] have adapted contrastive objectives to ViT backbones, demonstrating superior scalability. However, these methods typically operate on global image tokens, overlooking the multi-scale structural consistency required for fine-grained tasks.

III. Method

A. Overview

For the task of pollen image classification under electron microscopy, we propose a novel framework that integrates hierarchical background suppression with contrastive feature decoupling. As illustrated in Fig. 2, our model consists of three main components: (1) a vHeat backbone network that extracts multi-scale features through its hierarchical structure while modeling global long-range dependencies via heat conduction-based operators, (2) a Hierarchical Multi-scale Background Suppression (HMBS) module that progressively suppresses background interference across different feature scales, and (3) a Contrastive Feature Decoupling (CFD) module that explicitly separates foreground discriminative features from background noise through contrastive learning.

Given an input image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$, the vHeat backbone first extracts a pyramid of features $\{\mathbf{F}_i\}_{i=1}^{4}$, where $\mathbf{F}_i \in \mathbb{R}^{C_i\times H_i\times W_i}$ represents the feature map at stage i. The HMBS module then processes these multi-scale features through bidirectional information flows—a top-down semantic path and a bottom-up detail path—to generate importance weight maps $\{\mathbf{W}_i^{int}\}_{i=1}^{4}$ that identify discriminative foreground regions. Finally, the CFD module leverages these weight maps to extract foreground and background representations, applying a three-level contrastive loss to maximize inter-class separation while minimizing intra-class variation.

The overall training objective combines the classification loss with the contrastive loss:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \alpha\mathcal{L}_{cfd} + \beta\mathcal{L}_{reg}, \tag{1}$$

where $\mathcal{L}_{cls}$ is the cross-entropy classification loss, $\mathcal{L}_{cfd}$ is the multi-level contrastive loss from CFD, $\mathcal{L}_{reg}$ represents regularization terms including uncertainty regularization and foreground ratio regularization, and α, β are balancing coefficients.

B. vHeat Backbone

We adopt vHeat [11] for efficient global modeling via heat-diffusion-inspired operators. Following [11], vHeat outputs four-stage multi-scale features $\{\mathbf{F}_i\}_{i=1}^{4}$ for downstream suppression and decoupling.

C. Hierarchical Multi-scale Background Suppression (HMBS)

Although vHeat provides efficient global modeling, its multi-scale features still exhibit strong background responses under electron microscopy (Fig. 1). HMBS addresses this by producing stage-wise *foreground importance maps* and using them to suppress task-irrelevant activations. Specifically, HMBS aggregates complementary cues via a bidirectional fusion (top-down semantics and bottom-up details), followed by lightweight refinement, and learns a gate to balance the two streams.

For each stage i, HMBS outputs an importance map $\mathbf{W}_i^{int}$ used for subsequent masking:

$$\mathbf{W}_i^{int} = \boldsymbol{\gamma}_i \odot \mathbf{w}_i^{sem} + (1 - \boldsymbol{\gamma}_i) \odot \mathbf{w}_i^{det}, \tag{2}$$

where $\mathbf{w}_i^{sem}$ and $\mathbf{w}_i^{det}$ denote the semantic/detail cues and $\boldsymbol{\gamma}_i \in [0, 1]$ is a learnable gate.

1) Uncertainty-Aware Adaptive Thresholding (UAAT): Although the fused weight map $\mathbf{W}_i^{int}$ highlights potential foreground regions, converting these continuous activations into effective suppression masks is non-trivial. Simply applying a fixed threshold across all images ignores the varying levels of ambiguity in electron microscopy data. To address this, we propose **Uncertainty-Aware Adaptive Thresholding (UAAT)**, which dynamically calibrates the suppression threshold based on both image content and predictive uncertainty.

The threshold prediction network consists of a global average pooling layer followed by a two-layer MLP. During training, we employ Monte Carlo (MC) Dropout to estimate prediction uncertainty. Specifically, a spatial dropout layer with rate $p = 0.1$ is applied to the refined features $\mathbf{R}_i$ before the threshold prediction network. We perform $N = 3$ stochastic forward passes to obtain threshold samples:

$$\tau_i^{(n)} = \text{ThresholdNet}(\text{Dropout}(\mathbf{R}_i)), \quad n = 1, \ldots, N. \tag{3}$$

We then compute the mean and variance of these predictions:

$$\bar{\tau}_i = \frac{1}{N} \sum_{n=1}^{N} \tau_i^{(n)}, \quad \sigma_{\tau,i}^2 = \text{Var}\left(\{\tau_i^{(n)}\}_{n=1}^{N}\right). \tag{4}$$

The mean $\bar{\tau}_i$ serves as the content-dependent threshold tailored to the current image, while the variance $\sigma_{\tau,i}^2$ quantifies prediction uncertainty. The final suppression threshold is adaptively adjusted as:

$$\tau_{\text{adapt},i} = \text{clamp}\left(\bar{\tau}_i - \kappa \cdot \sigma_{\tau,i}^2,\ \tau_{\min},\ \tau_{\max}\right), \tag{5}$$

where κ is a learnable scaling factor, and $[\tau_{\min}, \tau_{\max}]$ defines the valid threshold range. This mechanism lowers the threshold when predictions exhibit high uncertainty (large σ^2), thereby conservatively preserving potential foregrounds.

During inference, we disable MC sampling and predict the threshold deterministically:

$$\tau_{\text{adapt},i} = \text{clamp}\left(\text{ThresholdNet}(\mathbf{R}_i),\ \tau_{\min},\ \tau_{\max}\right). \tag{6}$$

Finally, soft thresholding with a learnable temperature T is applied:

$$\mathbf{M}_i^{\text{soft}} = \sigma\left((\mathbf{W}_i^{int} - \tau_{\text{adapt},i}) \cdot T\right), \tag{7}$$

where $\sigma(\cdot)$ denotes the sigmoid function.

D. Contrastive Feature Decoupling (CFD)

Although HMBS suppresses dominant background activations, residual interference may remain, and semantic consistency across scales is not explicitly enforced. We therefore propose **CFD**, which (i) discovers foreground regions via a differentiable decision and (ii) improves discriminability with foreground–background contrast and cross-scale consistency.

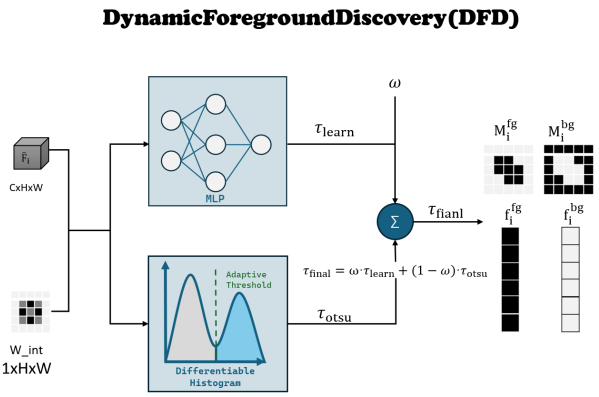

Fig. 3. Dynamic Foreground Discovery (DFD) module architecture.

1) Dynamic Foreground Discovery (DFD): As illustrated in Fig. 3, given $\hat{\mathbf{F}}_i$ and $\mathbf{W}_i^{int}$ at stage i, DFD estimates a robust threshold by fusing a learned predictor with a differentiable Otsu criterion:

$$\tau_i^{final} = \omega \cdot \tau_i^{learn} + (1 - \omega) \cdot \tau_i^{otsu}, \quad \omega = \sigma(w). \tag{8}$$

Foreground/background masks are obtained by soft thresholding:

$$\mathbf{M}_i^{fg} = \sigma((\mathbf{W}_i^{int} - \tau_i^{final}) \cdot T), \quad \mathbf{M}_i^{bg} = 1 - \mathbf{M}_i^{fg}. \tag{9}$$

We then extract stage-wise foreground/background features via masked average pooling and concatenate them across stages to form global representations $\mathbf{F}^{fg}$ and $\mathbf{F}^{bg}$.

2) Multi-level Contrastive Learning: Given the extracted foreground and background features, CFD further enhances foreground discriminability through multi-level contrastive learning. We design a three-level framework consisting of local, global, and cross-sample contrastive objectives.

Local contrastive loss operates at each individual scale, pulling together same-class foreground features while pushing apart foreground and background features. We formulate it as:

$$\mathcal{L}_{local}^{(i)} = -\frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \log \frac{e^{(\mathbf{f}_i^{fg} \cdot \mathbf{f}_p^{fg}/\tau)}}{e^{(\mathbf{f}_i^{fg} \cdot \mathbf{f}_p^{fg}/\tau)} + \sum_{n \in \mathcal{N}} e^{(\mathbf{f}_i^{fg} \cdot \mathbf{f}_n/\tau)}}, \tag{10}$$

where $\mathcal{P}$ is the set of same-class foregrounds. Crucially, the negative set $\mathcal{N}$ explicitly includes the current stage background $\mathbf{f}_i^{bg}$ alongside other samples' backgrounds, strictly enforcing foreground-background separability. The overall local loss is $\mathcal{L}_{local} = \frac{1}{4} \sum_{i=1}^{4} \mathcal{L}_{local}^{(i)}$.

Global contrastive loss is computed on the aggregated representations $\mathbf{F}^{fg}$ and $\mathbf{F}^{bg}$. Following the same logic, it enforces class-level discrimination at a higher semantic level:

$$\mathcal{L}_{global} = \mathcal{L}_{NCE}(\mathbf{F}^{fg}, \mathcal{P}, \{\mathbf{F}^{bg}\} \cup \mathcal{N}_{batch}), \tag{11}$$

where $\mathcal{L}_{NCE}$ denotes the operator in Eq. (10), and $\mathbf{F}^{bg}$ serves as the hardest negative to prevent global feature drift.

Cross-sample contrastive loss employs hard negative mining to improve robustness against challenging cases. We select the top-k most similar background features from the batch $\mathcal{B}_{bg}$ to form the hard negative set $\mathcal{N}_{hard}$:

$$\mathcal{N}_{hard} = \text{Top-}k(\{\mathbf{F}^{fg} \cdot \mathbf{b} \mid \mathbf{b} \in \mathcal{B}_{bg}\}). \tag{12}$$

The loss is then defined as $\mathcal{L}_{sample} = \mathcal{L}_{NCE}(\mathbf{F}^{fg}, \mathcal{P}, \mathcal{N}_{hard})$, forcing the model to distinguish the object from confusing textures in other images.

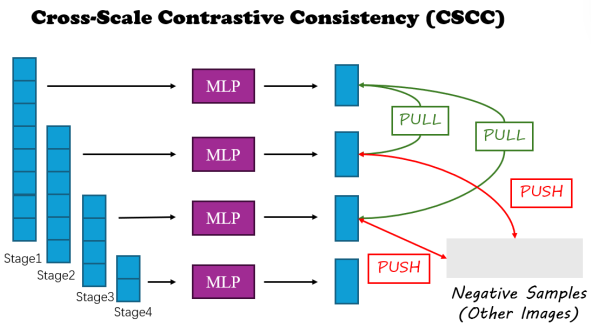

Fig. 4. Cross-Scale Contrastive Consistency (CSCC).

3) Cross-Scale Contrastive Consistency (CSCC): As shown in Fig. 4, to prevent scale-specific drifts, CSCC projects stage-wise foreground features into a shared space:

$$\mathbf{z}_i = \text{Proj}_i(\mathbf{f}_i^{fg}), \quad \mathbf{z}_i \in \mathbb{R}^d. \tag{13}$$

We enforce both adjacent and long-range cross-scale consistency:

$$\mathcal{L}_{cscc} = \sum_{i=1}^{3} \mathcal{L}_{align}(\mathbf{z}_i, \mathbf{z}_{i+1}) + \mathcal{L}_{align}(\mathbf{z}_1, \mathbf{z}_3) + \mathcal{L}_{align}(\mathbf{z}_2, \mathbf{z}_4), \tag{14}$$

where $\mathcal{L}_{align}(\mathbf{z}_i, \mathbf{z}_j)$ is an InfoNCE-style alignment loss. For each sample, $(\mathbf{z}_i, \mathbf{z}_j)$ from the **same image** forms the positive pair, while cross-scale features from **different-class samples** in the batch serve as negatives. Same-class samples are excluded from the negative set to preserve intra-class coherence.

4) Overall CFD Objective:

$$\mathcal{L}_{cfd} = \mathcal{L}_{local} + \mathcal{L}_{global} + \mathcal{L}_{sample} + \lambda_{cscc}\mathcal{L}_{cscc}. \tag{15}$$

IV. Experiments

A. Datasets and Experimental Setup

To comprehensively evaluate the effectiveness and generalization capability of the proposed method, we conduct experiments on four pollen datasets spanning diverse geographical regions, including Asia, Europe, Oceania, and South America.

We construct a large-scale pollen dataset with samples collected from Hohhot, Inner Mongolia, China, between 2018 and 2024, covering 51 families and 149 species. All images are captured using scanning electron microscopy (SEM), resulting in 29,018 images with a resolution of 224×224 pixels. The primary challenge of this dataset lies in its significant intra-class morphological variation coupled with high inter-class similarity, particularly evident in the Asteraceae family, which encompasses 16 morphologically similar genera including *Taraxacum*, *Helianthus*, and *Artemisia*. The dataset is partitioned into training, validation, and test sets at a ratio of 4:1:1.

Additionally, three publicly available benchmarks are employed to validate cross-domain generalization. POLLEN23E comprises 790 images across 23 European pollen categories. The New Zealand Pollen Dataset contains 46 categories totaling 19,667 darkfield microscopy images with substantial class imbalance. POLLEN73S originates from the Brazilian Cerrado ecosystem, containing 2,523 multi-view images across 73 categories.

Our framework is developed with PyTorch 2.1.0 on a single NVIDIA RTX 3090 GPU. To ensure fair comparison, all experiments are conducted under identical environmental settings. The backbone network is vHeat and is trained from scratch with random initialization, without using any pretrained weights. Training is performed for 300 epochs with a batch size of 128. We use AdamW optimizer with an initial learning rate of 5×10^{-4}, cosine decay schedule, and weight decay of 0.05. Augmentation strategies include RandAugment, Mixup, and CutMix.

B. Comparison with State-of-the-Art Methods

We compare our proposed framework with several state-of-the-art methods, including standard CNNs (ResNet [6], ConvNeXt [15]), Vision Transformers (ViT [7], Swin Transformer [8]), and recent advanced models (VSSD [36], MogaNet [18]). The quantitative results are summarized in Table I.

On our self-constructed electron microscopy dataset, **Pollen149**, our method achieves a Top-1 accuracy of **98.1%**, significantly outperforming the baseline vHeat and other competing models. This demonstrates the superiority of our approach in handling fine-grained pollen morphological details. Furthermore, on the three public benchmarks, our model consistently exhibits excellent generalization capability, achieving state-of-the-art performance across diverse data distributions.

C. Ablation Study

To validate the effectiveness of the proposed modules and their sub-modules, we conduct ablation experiments on the

TABLE I
COMPARISON WITH STATE-OF-THE-ART METHODS ON FOUR POLLEN DATASETS (%)

Method	Pollen149 (Ours)				POLLEN73S				POLLEN23E				New Zealand			
	ACC	**Prec**	**Rec**	**F1**	**ACC**	**Prec**	**Rec**	**F1**	**ACC**	**Prec**	**Rec**	**F1**	**ACC**	**Prec**	**Rec**	**F1**
ResNet	87.3	87.8	85.1	85.2	80.4	81.0	80.0	79.1	68.4	68.7	68.5	67.0	96.8	96.2	95.4	95.7
ViT	73.6	72.2	69.8	69.6	81.0	83.0	81.1	80.6	74.7	78.6	75.2	74.1	91.0	90.8	88.7	89.5
Swin Transformer	95.6	95.5	94.1	94.5	84.1	83.1	82.0	81.6	72.8	76.8	73.5	72.8	97.1	96.6	95.7	96.0
ConvNeXt	86.2	86.6	83.2	83.7	73.7	74.5	73.8	72.6	75.3	77.9	75.3	74.8	95.3	94.3	93.5	93.8
VSSD	91.8	90.7	89.7	89.7	79.6	79.6	78.5	79.2	75.3	77.9	75.3	74.8	97.1	96.3	95.8	95.5
MogaNet	85.1	86.3	81.6	82.0	78.5	82.0	79.4	77.4	75.9	79.9	77.2	75.4	98.2	97.9	97.2	97.2
vHeat	96.3	96.0	94.5	94.9	88.5	89.9	88.1	88.7	82.3	84.6	81.9	80.3	98.6	98.1	97.3	97.3
Ours	**98.1**	**97.7**	**97.4**	**97.3**	**90.5**	**91.8**	**90.6**	**90.1**	**83.5**	**85.7**	**83.7**	**82.1**	**98.9**	**98.5**	**98.1**	**98.1**

TABLE II
ABLATION STUDY ON THE SELF-CONSTRUCTED DATASET

Configuration	ACC (%)
Baseline(vHeat)	96.31
HMBS	97.86
HMBS w/o UAAT	97.50
CFD (Full)	97.84
CFD w/o CSCC	97.73
CFD w/o CSCC & DFD	97.30
Full Model (HMBS + CFD)	**98.08**

TABLE III
PARAMETER SENSITIVITY ANALYSIS ON THE SELF-CONSTRUCTED DATASET

Parameter	Value	ACC (%)
Contrastive weight α	0	97.50
	0.05	**98.08**
	0.1	97.80
	0.2	97.80
	0.5	97.52
Temperature τ	0.03	98.01
	0.05	**98.08**
	0.07	97.91
	0.1	98.01
	0.15	97.95
MC samples	1	97.69
	3	**98.08**
	5	97.52
CSCC weight λ	0	97.79
	0.1	**98.08**
	0.5	97.52

self-constructed dataset. The results are summarized in Table II.

In the background suppression branch, the complete HMBS module achieves an accuracy of 97.86%. Removing the UAAT (Uncertainty-Aware Adaptive Thresholding) sub-module decreases the accuracy to 97.50%, a drop of 0.36 percentage points. This demonstrates that UAAT effectively enhances foreground boundary localization through uncertainty modeling.

In the feature decoupling branch, the complete CFD module achieves an accuracy of 97.84%. When both CSCC and DFD modules are removed, the accuracy drops significantly to 97.30%, a substantial decline of 0.54% that underscores the foundational role of the DFD mechanism. Adding DFD back (i.e., removing only CSCC) recovers the accuracy to 97.73%. The inclusion of CSCC further improves the performance to 97.84%. While this marginal gain might appear modest, it is noteworthy given the high baseline performance (>97.7%), suggesting that enforcing cross-scale consistency provides a complementary regularization effect that helps resolve ambiguous cases.

Finally, the full model integrating both HMBS and CFD branches achieves the highest accuracy of 98.08%, outperforming individual branches by 0.22% and 0.24%, respectively. This fully demonstrates the synergistic effect of the two branch modules.

D. Parameter Sensitivity Analysis

To investigate the impact of key hyperparameters on model performance, we conduct parameter sensitivity analysis on the self-constructed dataset. The results are summarized in Table III.

Contrastive loss weight α: This parameter controls the contribution of contrastive learning. Setting $\alpha = 0$ (no contrastive learning) results in 97.50% accuracy. Introducing a small weight of $\alpha = 0.05$ significantly boosts performance to the peak of **98.08%** (+0.58%), validating the necessity of the contrastive branch. Further increasing α leads to a slight performance drop, suggesting that the auxiliary loss should not dominate the primary classification objective.

Temperature coefficient τ: This parameter regulates the feature distribution smoothness. The model exhibits high robustness to τ, maintaining accuracy above 97.9% across a wide range. However, fine-tuning τ to 0.05 yields the optimal accuracy of 98.08%, ensuring the best trade-off between class compactness and separability.

MC sampling times: The model performance is relatively stable with respect to sampling frequency. While a single sample achieves 97.69%, increasing it to 3 brings the accuracy to 98.08%. This improvement indicates that appropriate uncertainty estimation helps, but excessive sampling (e.g., 5) is unnecessary and may introduce noise.

CSCC loss weight λ: Comparing $\lambda = 0$ (97.79%) with $\lambda = 0.1$ (98.08%), we observe a clear gain of 0.29% from the cross-scale consistency constraint. The performance remains robust for smaller weights but degrades when λ is too large (0.5), confirming that CSCC serves as a regularization term rather than a primary objective.

In summary, while the model is generally robust to hyperparameter variations, appropriate tuning allows it to reach its

performance ceiling. We adopt $\alpha = 0.05$, $\tau = 0.05$, MC samples$= 3$, and $\lambda = 0.1$ as the default configuration.

V. Conclusion

In this paper, we presented a method for electron microscopy pollen image classification designed to address the dual challenges of complex background interference and high inter-class morphological similarity. We introduced the Hierarchical Multi-scale Background Suppression (HMBS) module, which leverages Uncertainty-Aware Adaptive Thresholding (UAAT) to dynamically filter task-irrelevant noise while preserving critical pollen structures. Furthermore, the Contrastive Feature Decoupling (CFD) module was proposed to enhance feature discriminability through Dynamic Foreground Discovery (DFD) and Cross-Scale Contrastive Consistency (CSCC). Extensive experiments on our self-constructed large-scale Pollen149 dataset and three public benchmarks demonstrate that our method achieves superior performance, reaching a top-1 accuracy of 98.1% on Pollen149. These results confirm the effectiveness of our approach in automated pollen classification, offering a promising solution for biodiversity monitoring and paleoecological research. Despite these promising results, one limitation is that the multi-scale feature fusion introduces additional computational overhead. In future work, we plan to explore more lightweight architectural designs to further improve both accuracy and inference efficiency.

Acknowledgment

This work was supported by Science and Technology Plan Project of Inner Mongolia Autonomous Region (2025YFHH0083, 2025YFHH0215), Inner Mongolia Natural Science Foundation Project (2025MS06039, 2023LHMS07016, 2025QN06005), Aerospace Medical Health Technology Group Co., Ltd. Medical and Industrial Integration Project (2025YK13), and the Inner Mongolia Science and Technology Innovation Guidance Project (CXYD2022BT07).

References

[1] J. Flenley, "Tropical forests under the climates of the last 30,000 years," *Climatic change*, vol. 39, no. 2, pp. 177–197, 1998.

[2] G. D'Amato, C. Vitale, A. Sanduzzi, A. Molino, A. Vatrella, and M. D'Amato, "Allergenic pollen and pollen allergy in europe," *Allergy and allergen immunotherapy*, pp. 287–306, 2017.

[3] D. Mildenhall, "Hypericum pollen determines the presence of burglars at the scene of a crime: an example of forensic palynology," *Forensic Science International*, vol. 163, no. 3, pp. 231–235, 2006.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[5] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2021.

[8] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[9] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang, "Vision mamba: Efficient visual representation learning with bidirectional state space model," *arXiv preprint arXiv:2401.09417*, 2024.

[10] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, J. Jiao, and Y. Liu, "Vmamba: Visual state space model," *Advances in neural information processing systems*, vol. 37, pp. 103 031–103 063, 2024.

[11] Z. Wang, Y. Liu, Y. Tian, Y. Liu, Y. Wang, and Q. Ye, "Building vision models upon heat conduction," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 9707–9717.

[12] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[13] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[14] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[15] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 976–11 986.

[16] X. Ding, X. Zhang, N. Ma, J. Han, G. Ding, and J. Sun, "Repvgg: Making vgg-style convnets great again," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 13 733–13 742.

[17] X. Ding, X. Zhang, J. Han, and G. Ding, "Scaling up your kernels to 31x31: Revisiting large kernel design in cnns," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 963–11 975.

[18] S. Li, Z. Wang, Z. Liu, C. Tan, H. Lin, D. Wu, Z. Chen, J. Zheng, and S. Z. Li, "Moganet: Multi-order gated aggregation network," *arXiv preprint arXiv:2211.03295*, 2022.

[19] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[20] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6848–6856.

[21] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 116–131.

[22] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 764–773.

[23] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable convnets v2: More deformable, better results," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 9308–9316.

[24] J. Li, Y. Wen, and L. He, "Scconv: Spatial and channel reconstruction convolution for feature redundancy," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 6153–6162.

[25] Y. Xiong, Z. Li, Y. Chen, F. Wang, X. Zhu, J. Luo, W. Wang, T. Lu, H. Li, Y. Qiao *et al.*, "Efficient deformable convnets: Rethinking dynamic and sparse operator for vision applications," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 5652–5661.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[27] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 568–578.

[28] ——, "Pvt v2: Improved baselines with pyramid vision transformer," *Computational visual media*, vol. 8, no. 3, pp. 415–424, 2022.

[29] X. Chu, Z. Tian, Y. Wang, B. Zhang, H. Ren, X. Wei, H. Xia, and C. Shen, "Twins: Revisiting the design of spatial attention in vision transformers," *Advances in neural information processing systems*, vol. 34, pp. 9355–9366, 2021.

[30] X. Liu, H. Peng, N. Zheng, Y. Yang, H. Hu, and Y. Yuan, "Efficientvit: Memory efficient vision transformer with cascaded group attention,"

in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 14 420–14 430.

[31] X. Chen, Z. Liu, H. Tang, L. Yi, H. Zhao, and S. Han, "Sparsevit: Revisiting activation sparsity for efficient high-resolution vision transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2061–2070.

[32] C. Wei, B. Duke, R. Jiang, P. Aarabi, G. W. Taylor, and F. Shkurti, "Sparsifiner: Learning sparse instance-dependent attention for efficient vision transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 680–22 689.

[33] Z. Wu, T. Ding, Y. Lu, D. Pai, J. Zhang, W. Wang, Y. Yu, Y. Ma, and B. D. Haeffele, "Token statistics transformer: Linear-time attention via variational rate reduction," *arXiv preprint arXiv:2412.17810*, 2024.

[34] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," in *First conference on language modeling*, 2024.

[35] L. Liu, M. Zhang, J. Yin, T. Liu, W. Ji, Y. Piao, and H. Lu, "Defmamba: Deformable visual state space model," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 8838–8847.

[36] Y. Shi, M. Li, M. Dong, and C. Xu, "Vssd: Vision mamba with non-causal state space duality," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 10 819–10 829.

[37] P. Viertel and M. König, "Pattern recognition methodologies for pollen grain image classification: a survey," *Machine Vision and Applications*, vol. 33, no. 1, p. 18, 2022.

[38] M. A. Rostami, B. Balmaki, L. A. Dyer, J. M. Allen, M. F. Sallam, and F. Frontalini, "Efficient pollen grain classification using pre-trained convolutional neural networks: a comprehensive study," *Journal of big data*, vol. 10, no. 1, p. 151, 2023.

[39] G. Astolfi, A. B. Goncalves, G. V. Menezes, F. S. B. Borges, A. C. M. N. Astolfi, E. T. Matsubara, M. Alvarez, and H. Pistori, "Pollen73s: An image dataset for pollen grains classification," *Ecological Informatics*, vol. 60, p. 101165, 2020.

[40] V. Sevillano and J. L. Aznarte, "Improving classification of pollen grain images of the polen23e dataset through three different applications of deep learning convolutional neural networks," *PloS one*, vol. 13, no. 9, p. e0201807, 2018.

[41] M. Zolfaghari and H. Sajedi, "Automated classification of pollen grains microscopic images using cognitive attention based on human two visual streams hypothesis," *PLoS One*, vol. 19, no. 11, p. e0309674, 2024.

[42] T. Mahmood, J. Choi, and K. R. Park, "Artificial intelligence-based classification of pollen grains using attention-guided pollen features aggregation network," *Journal of King Saud University-Computer and Information Sciences*, vol. 35, no. 2, pp. 740–756, 2023.

[43] Y. Long, W. Sun, N. Sun, W. Wang, C. Li, and S. Yin, "Hieraedgenet: A multi-scale edge-enhanced framework for automated pollen recognition," *Agriculture*, vol. 15, no. 23, p. 2518, 2025.

[44] B. Zu, T. Cao, Y. Li, J. Li, H. Wang, and Q. Wang, "Reswint: enhanced pollen image classification with parallel window transformer and coordinate attention," *The Visual Computer*, vol. 41, no. 7, pp. 4975–4990, 2025.

[45] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PmLR, 2020, pp. 1597–1607.

[46] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, "Big self-supervised models are strong semi-supervised learners," *Advances in neural information processing systems*, vol. 33, pp. 22 243–22 255, 2020.

[47] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[48] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," *arXiv preprint arXiv:2003.04297*, 2020.

[49] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in neural information processing systems*, vol. 33, pp. 21 271–21 284, 2020.

[50] X. Chen and K. He, "Exploring simple siamese representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 15 750–15 758.

[51] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," *Advances in neural information processing systems*, vol. 33, pp. 9912–9924, 2020.

[52] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9640–9649.

[53] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660.