Cross-Domain Tomato Disease Classification via Flexible Contrastive Clustering in Vision-Language Models

Muhammad Shafay

100057573@ku.ac.ae

Khalifa University, Abu Dhabi, UAE

Divya Velayudhan

Khalifa University, Abu Dhabi, UAE

Taimur Hassan

Abu Dhabi University, Abu Dhabi, UAE

Muhammad Owais

Khalifa University, Abu Dhabi, UAE

Irfan Hussain

Khalifa University, Abu Dhabi, UAE

Naoufel Werghi

Khalifa University, Abu Dhabi, UAE

Research Article

Keywords: Vision-language models, cross-domain generalization, plant disease classification, flexible contrastive clustering, agricultural computer vision

Posted Date: March 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9198236/v1

Additional Declarations: The authors declare no competing interests.

Cross-Domain Tomato Disease Classification via Flexible Contrastive Clustering in Vision-Language Models

Muhammad Shafay*, Divya Velayudhan*, Taimur Hassan†, Muhammad Owais‡, Irfan Hussain‡, Naoufel Werghi*
*Department of Computer Science, Khalifa University, Abu Dhabi, UAE
Email: {100057573, divya.velayudhan, naoufel.werghi}@ku.ac.ae
†Department of Electrical & Computer Engineering, Abu Dhabi University, Abu Dhabi, UAE
Email: taimur.hassan@adu.ac.ae
‡Department of Mechanical Engineering, Khalifa University, Abu Dhabi, UAE
Email: {muhammad.owais, irfan.hussain}@ku.ac.ae

***Abstract*—Plant disease detection systems face significant challenges in cross-domain generalization, particularly when transitioning from controlled laboratory settings to diverse field conditions. Traditional deep learning approaches exhibit severe performance degradation across different imaging environments, limiting practical deployment in real-world agricultural scenarios. This paper introduces a novel Flexible Contrastive Clustering (FCC) framework for zero-shot tomato disease classification that addresses fundamental generalization limitations through vision-language learning. Unlike standard CLIP's one-to-one image-text pairing, our method leverages one-to-many relationships where each disease image is associated with multiple diverse textual descriptions, enabling robust representation learning across linguistic variations. The FCC framework optimizes class-based clustering in joint embedding space through a specialized loss function that treats all same-class descriptions as positives, facilitating effective handling of both seen and unseen disease categories during zero-shot evaluation. We evaluate our approach on PlantDoc training data (740 images) and test across four diverse tomato disease datasets totaling 17,313 images, spanning laboratory and field conditions. Experimental results demonstrate substantial improvements over state-of-the-art vision-language models, achieving an average of 30.15% accuracy and 28.05% weighted F1-score on average across all test datasets. Our method shows particularly strong performance on field datasets, achieving 59.70% accuracy on FieldPlant and 26.52% on Tomato Village, significantly outperforming existing approaches. Attention visualization analysis reveals effective disease localization capabilities for both seen and unseen categories, validating the practical applicability of our approach for real-world agricultural monitoring systems.**

***Index Terms*—Vision-language models, cross-domain generalization, plant disease classification, flexible contrastive clustering, agricultural computer vision**

I. Introduction

Tomato production, the world's second most important vegetable crop after potato, faces critical challenges threatening global food security. With approximately 186 million tonnes produced in 2022 [1], pathogen-induced diseases cause 10-50% yield losses in typical scenarios, with severe outbreaks reaching 80% reduction. For smallholder farmers comprising over 80% of global agricultural producers, disease-related losses average 50% of total production, necessitating accessible diagnostic solutions.

Current diagnostic approaches rely on manual visual inspection by trained pathologists, suffering from subjective interpretation, high costs, limited expert availability, and human error susceptibility [2]. While Convolutional Neural Networks achieve over 95% accuracy on controlled datasets [3]–[5], fundamental generalization limitations emerge during real-world deployment. Laboratory datasets such as PlantVillage feature controlled conditions with isolated specimens against high-contrast backgrounds, whereas field environments present heterogeneous visual contexts where pathological symptoms exhibit reduced discriminability due to spectral overlap with healthy vegetation and complex background interference.

The primary barrier to practical deployment is poor cross-domain generalization. Traditional models [6]–[9] exhibit severe performance degradation when applied to conditions differing from training environments [9]–[14] overfitting to dataset-specific characteristics rather than learning generalizable disease features. Figure 1 illustrates this challenge across five tomato disease datasets, revealing stark differences between field-captured datasets (natural lighting, complex backgrounds, diverse orientations) and laboratory-controlled datasets (uniform backgrounds, standardized protocols).

Vision-Language Models (VLMs) offer transformative potential for addressing generalization challenges through large-scale joint visual-textual pretraining. CLIP demonstrates remarkable zero-shot capabilities by learning from 400 million image-text pairs, achieving zero-shot ImageNet performance comparable to supervised ResNet-50 [15]. Unlike supervised approaches, CLIP develops general visual representations aligned with natural language, enabling transfer without task-specific training. However, direct application of general-purpose VLMs to specialized agricultural tasks reveals performance gaps due to insufficient fine-grained discriminative capabilities for subtle disease symptom detection.

This paper introduces a novel vision-language approach for cross-dataset tomato disease classification addressing fun-

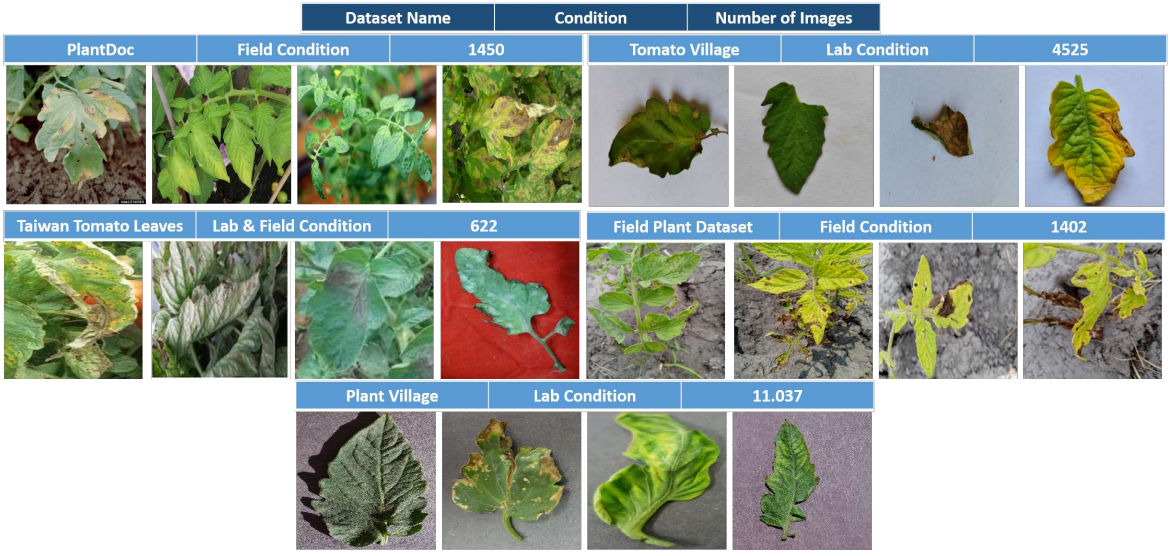

Fig. 1: Comparative analysis of tomato disease datasets demonstrating diverse imaging conditions. Field datasets (PlantDoc, Tomato Leaf Dataset, Field Plant Dataset) exhibit natural lighting and complex backgrounds, while laboratory datasets (Plant Village, Tomato Village, Taiwan Tomato Leaves Dataset) feature controlled environments with standardized protocols.

damental generalization limitations. Our Flexible Contrastive Clustering (FCC) framework optimizes class-based clustering in joint vision-language embedding space for enhanced cross-domain transferability. Key contributions include:

- A Flexible Contrastive Clustering framework optimizing class-based clustering in joint vision-language embedding space for effective handling of seen and unseen disease categories during zero-shot evaluation.
- A dynamic training approach that leverages multiple textual descriptions per image during contrastive learning, enhancing semantic diversity without requiring additional data collection.
- Comprehensive zero-shot evaluation across four diverse tomato disease datasets, establishing new cross-dataset plant disease classification benchmarks.
- Interpretability through attention visualization, revealing effective disease localization for both seen and unseen categories, with substantial improvements over state-of-the-art vision-language models.

II. Related Work

A. Deep Learning for Plant Disease Detection

Deep learning [16] has revolutionized plant disease detection capabilities. Mohanty et al. [17] pioneered CNN application to plant disease classification using PlantVillage, achieving 99.35% accuracy across 14 crop species and 26 diseases under controlled conditions. Khan et al. [18] advanced this field by introducing convolutional transformers for tomato maturity recognition, demonstrating the potential of hybrid architectures that combine CNN feature extraction with transformer attention mechanisms. Subsequent works include Agarwal et al. [19] developing ToLeD with custom CNN architectures (91.2% tomato disease accuracy) and Fuentes et al. [20] applying Faster R-CNN for real-time field detection (96% success rate).

However, these models exhibit poor cross-dataset generalization. Recent reviews [21], [22] highlight significant performance degradation when models trained on controlled environments are tested on uncontrolled field conditions [23].

B. Vision-Language Models and Cross-Domain Adaptation

CLIP [15] marked a paradigm shift, enabling zero-shot classification through joint vision-language representation learning with ImageNet performance comparable to supervised ResNet-50. However, CLIP's application to specialized domains remains limited due to its one-to-one alignment paradigm treating each image-text pair independently.

Vision-language models show promise in medical imaging with similar challenges. Wang et al. [24] demonstrated that MedCLIP's multiple image-text relationships achieve superior generalization over standard CLIP, while Li et al. [25] showed that local image-text matching relations significantly improve representation learning. However, set-based contrastive learning approaches remain unexplored in agricultural disease detection.

Cross-domain generalization represents a critical challenge in agricultural AI deployment. Studies document the laboratory-to-field gap [21], [26], with existing domain adaptation approaches primarily focusing on single-dataset optimization or requiring target domain data during training [27], [28].

C. Research Gap and Motivation

Despite advances in plant disease detection and vision-language models, critical gaps persist: poor cross-dataset generalization, particularly from laboratory to field conditions [29], [30], and CLIP's limited performance on specialized agricultural tasks due to its one-to-one alignment paradigm.

Our work addresses these gaps through a Flexible Contrastive Clustering framework leveraging one-to-many relationships between disease images and multiple semantic descriptions. This represents the first systematic evaluation of vision-language models for cross-dataset plant disease classification, establishing new agricultural computer vision benchmarks.

III. Methodology

A. CLIP Model Foundation

The Contrastive Language-Image Pre-training (CLIP) model learns joint representations of images and text through contrastive learning. CLIP consists of two core components: an image encoder $f(\cdot)$ that maps input images to embeddings, and a text encoder $g(\cdot)$ that encodes textual descriptions into the same embedding space.

Given an image x_i and its corresponding text t_i, CLIP generates embeddings:

$$\mathbf{v}_i = f(x_i) \quad \text{(image embedding)} \tag{1}$$

$$\mathbf{u}_i = g(t_i) \quad \text{(text embedding)} \tag{2}$$

CLIP optimizes a symmetric contrastive objective that maximizes cosine similarity between corresponding image-text pairs while minimizing similarity between non-corresponding pairs:

$$\mathscr{L}_{\text{CLIP}} = \frac{1}{2}(\mathscr{L}_{i2t} + \mathscr{L}_{t2i}) \tag{3}$$

where $\mathscr{L}_{i2t}$ and $\mathscr{L}_{t2i}$ represent image-to-text and text-to-image contrastive losses respectively, both computed using cosine similarity $\text{sim}(\mathbf{v}_i, \mathbf{u}_j) = \frac{\mathbf{v}_i \cdot \mathbf{u}_j}{|\mathbf{v}_i||\mathbf{u}_j|}$. For zero-shot classification, CLIP compares image embeddings with text embeddings generated from class-specific prompts to predict the most similar class.

B. Proposed Vision-Language Framework for Tomato Disease Classification

Our approach extends CLIP for tomato disease classification by leveraging multiple textual descriptions per image and incorporating class labels directly into learning. Unlike standard CLIP's one-to-one alignment, our framework processes multiple descriptions per image, enabling richer semantic understanding of disease characteristics. Figure 2 illustrates the proposed methodology.

Given training image x_i with disease class y_i, we associate it with multiple textual descriptions $\mathscr{D}_i = \{t_{i,1}, t_{i,2}, ..., t_{i,m}\}$ where m denotes descriptions per image. Each description contains visual symptom information and explicit class labels, facilitating direct label induction during training.

We constructed a dataset comprising 740 tomato disease images, each annotated with 5 diverse textual descriptions capturing different symptom characteristics. These descriptions encompass various visual aspects including lesion morphology, spatial distribution patterns, color variations, size descriptors, and disease-specific terminology to provide comprehensive linguistic coverage of pathological manifestations. For example, bacterial spot disease descriptions vary across size variations (*"small dark spots"* vs *"tiny black lesions"*), shape descriptors (*"round bacterial spots"*), and distribution patterns (*"scattered"* vs *"numerous"*). Similarly, early blight exhibits diverse linguistic representations ranging from *"brown circular lesions with concentric rings"* to *"dark spots with target-like patterns on leaf surface"*. This approach ensures that each disease class is represented through 5 linguistically diverse samples that capture complementary aspects of the visual phenotype. Through our dynamic training strategy, we establish cross-modal associations between images and their corresponding textual descriptions using a bag-of-words representation to encode semantic relationships within the joint embedding space. This process utilizes our foundational corpus of 3,700 image-text pairs (740×5), where each visual sample is paired with multiple linguistic representations from both original descriptions and visually similar samples. This multi-modal framework enables robust visual-linguistic associations, bridging the vocabulary gap between expert pathological descriptions and field-based observations.

C. Flexible Contrastive Clustering Loss

The Flexible Contrastive Clustering (FCC) loss aligns image and text embeddings by exploiting class-level supervision with multiple textual descriptions per image. This approach extends traditional one-to-one contrastive alignment by leveraging one-to-k correspondence between each image and its k diverse text descriptions, enhancing semantic robustness. In our experiments, k is set to 5.

Given batch B with images and $B_T = B \times k$ text descriptions, we compute normalized embeddings $f(x_i)$ for image x_i and $g(t_j)$ for text description t_j. Normalization ensures embeddings lie on a unit hypersphere for stable similarity comparisons. We calculate temperature-scaled similarity matrix $S \in \mathbb{R}^{B \times B_T}$ using cosine similarities:

$$S_{i,j} = \frac{f(x_i)^\top g(t_j)}{\tau}$$

where temperature τ controls distribution softness, enabling fine-grained similarity weighting. This matrix captures all potential image-text alignments, allowing consideration of diverse textual expressions and cross-image semantic relationships.

Class supervision is incorporated through binary mask matrix $M \in \{0,1\}^{B \times B_T}$:

$$M_{i,j} = \begin{cases} 1 & \text{if } x_i \text{ and } t_j \text{ share the same class} \\ 0 & \text{otherwise} \end{cases}$$

This mask treats all same-class text descriptions as positives, encouraging clustering of semantically similar images and texts while improving generalization across linguistic variations.

Similarity scores are transformed into log-probabilities via row-wise softmax:

$$\log P_{i,j} = S_{i,j} - \log \sum_{k=1}^{B_T} \exp(S_{i,k})$$

This probabilistic interpretation weights each textual description relative to others, enhancing discriminative capability by emphasizing correct class matches.

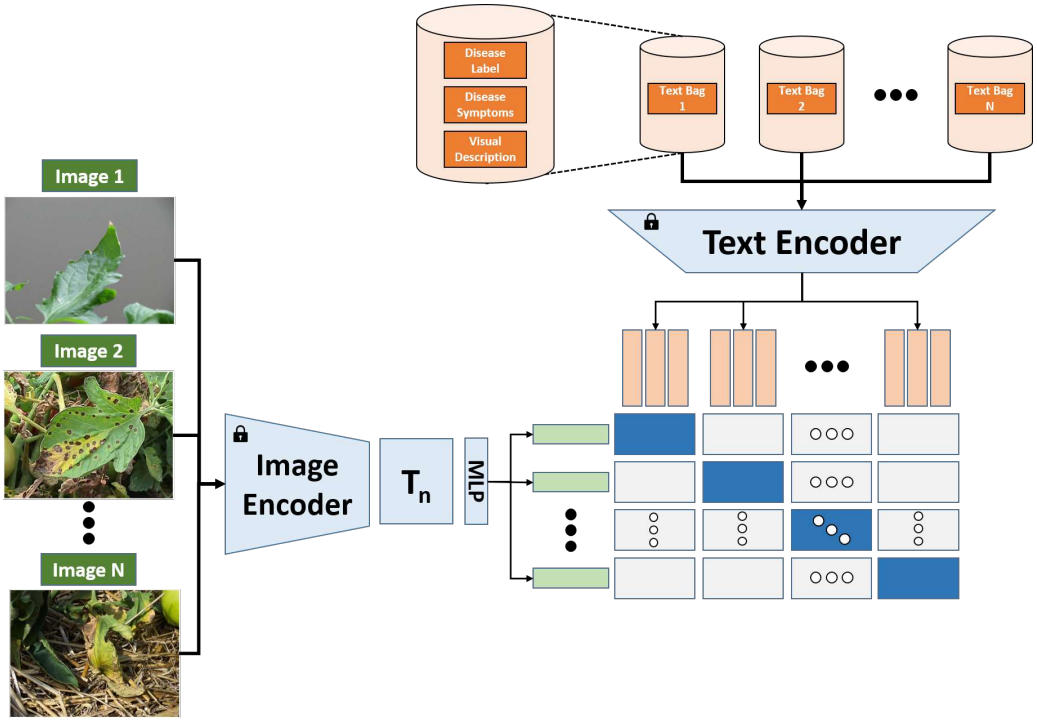

Fig. 2: Overview of the proposed Flexible Contrastive Clustering framework. Images are processed through a vision encoder while text bags containing disease descriptions are encoded through a frozen text encoder. During training, only the final transformer block (T_n) and MLP projection layer remain trainable in the vision encoder. The one-to-many image-text relationship enables richer semantic understanding compared to traditional CLIP.

The FCC loss averages log-probabilities over all positive matches defined by M:

$$\bar{P}_i = \frac{\sum_{j=1}^{B_T} M_{i,j} \cdot \log P_{i,j}}{\sum_{j=1}^{B_T} M_{i,j}}$$

This aggregation leverages multiple textual perspectives of the same class, producing smoother gradients and stable learning while exploiting intra-class diversity critical for agricultural disease recognition.

The final loss is computed as:

$$\mathscr{L}_{\text{FCC}} = -\frac{1}{B}\sum_{i=1}^{B} \bar{P}_i$$

By maximizing average probability of correctly matching images with related descriptions, FCC loss encourages class-wise clustering in joint embedding space, improving robustness against noisy annotations and enhancing cross-domain generalization.

D. Training and Zero-Shot Evaluation

The training process freezes the text encoder weights completely while selectively optimizing the vision encoder parameters. Specifically, in the vision encoder, only the final transformer block and MLP projection layer remain trainable, with all other layers frozen to preserve pre-trained visual representations. The model is optimized using the FCC loss with cosine annealing learning rate scheduling. Images undergo standard preprocessing including resizing to 224×224 pixels and normalization using ImageNet statistics before being processed through the vision encoder for multiple training epochs.

For zero-shot evaluation on unseen datasets, we employ multiple prompt templates to generate diverse textual descriptions for each disease class. The zero-shot classification score for an image and class is computed as the average similarity across all prompt variations. This multi-prompt evaluation strategy improves robustness and enables effective cross-dataset generalization.

TABLE I: Dataset Statistics for Training and Zero-Shot Evaluation for Tomato Leaf Disease Detection

Dataset	Images	Environment
Training Dataset		
PlantDoc [31]	740	Field
Zero-Shot Evaluation Datasets		
Taiwan Tomato Leaves Dataset [32]	622	Lab
PlantVillage [3]	11,037	Lab
FieldPlant [10]	1,402	Field
Tomato Village [33]	4,525	Field
Total Evaluation	**17,313**	

IV. Datasets and Experimental Setup

A. Dataset Description

Our experimental evaluation encompasses four diverse tomato disease datasets, strategically divided into training and

TABLE II: Performance Comparison Across Different Vision-Language Models

Dataset	Proposed		CLIP [15]		SigLIP [34]		LongCLIP [35]	
	Acc (%)	F1-W (%)	Acc (%)	F1-W (%)	Acc (%)	F1-W (%)	Acc (%)	F1-W (%)
PlantVillage [3]	23.80	20.30	21.21	21.73	9.54	2.91	**30.81**	**32.25**
Tomato Village [33]	**26.52**	**23.30**	20.88	11.53	1.94	0.74	21.77	13.55
Taiwan Tomato Leaves Dataset [32]	24.76	23.20	**26.69**	21.74	19.77	13.13	26.21	**22.50**
FieldPlant [10]	**59.70**	**63.40**	24.25	14.94	2.28	2.68	23.04	13.51
Average	**30.15**	**28.05**	19.99	15.20	15.00	8.96	23.13	18.25

evaluation sets to assess cross-dataset generalization capabilities across different imaging conditions and environments.

For model training, we utilize the PlantDoc dataset subset containing only tomato diseases [31], comprising 740 high-quality tomato leaf images. Each image is associated with multiple textual descriptions, providing diverse linguistic perspectives of the pathological conditions. Our approach leverages the one-to-multiple relationship between images and descriptions to enable rich semantic understanding.

Zero-shot evaluation is conducted across four independent datasets containing 17,313 images total, providing extensive coverage of different imaging conditions from controlled laboratory environments to challenging field conditions.

B. Experimental Configuration

The experimental setup utilizes the CLIP ViT-B/16 architecture as the foundation model. Training is conducted for 50 epochs with a batch size of 8 and learning rate of1×10^{-5}. The FCC loss temperature parameter is set to 0.07. During training, the text encoder weights remain frozen while in the image encoder, only the final transformer block and MLP projection layer are trainable, resulting in approximately 7.5 million trainable parameters. All experiments are performed on NVIDIA GPU hardware. For zero-shot evaluation, we employ multiple prompt templates to generate diverse textual descriptions for each disease class. The final classification decision is based on the highest average similarity score across all prompt variations.

V. Results and Discussion

We evaluate our Flexible Contrastive Clustering approach against three state-of-the-art vision-language models: CLIP, SigLIP, and LongCLIP. The model is trained on the PlantDoc dataset (740 images) and evaluated in a zero-shot manner across four diverse tomato disease datasets containing 17,313 images total. Table II presents comprehensive results including accuracy (Acc) and weighted F1-score (F1-W).

A. Overall Performance Analysis

Our proposed method achieves the highest average performance across all evaluation metrics, with 30.15% accuracy and 28.05% weighted F1-score, representing substantial improvements over baseline methods. Table II presents the comprehensive performance comparison across different vision-language models evaluated on four diverse tomato disease datasets totaling 17,313 images. Our method demonstrates particularly strong performance on field-captured datasets, achieving exceptional results on FieldPlant with 59.70% accuracy and 63.40% weighted F1-score compared to CLIP's 24.25% accuracy and 14.94% weighted F1-score. Similarly, on Tomato Village, we achieve 26.52% accuracy versus CLIP's 20.88%. These results highlight our method's effectiveness in real-world agricultural scenarios where imaging conditions are more challenging, with our approach showing consistent improvements over state-of-the-art vision-language models across both laboratory and field environments.

B. Attention Mechanism Visualization

Figure 3 presents attention heatmaps showing our model's focus on disease-relevant regions across both seen (S) and unseen (U) disease classes. For seen classes like Leaf Mold and Early Blight, attention maps demonstrate precise localization of disease symptoms with high attention weights (red/yellow regions) on areas encountered during training. More importantly, for unseen classes including Leaf Miner, Brown Spot, Black Mold, and Gray Spot, the model successfully identifies disease regions despite never seeing these diseases during training. This demonstrates effective generalization of disease-relevant features to novel categories, which is valuable for agricultural applications where new diseases may emerge.

The visualizations show consistent disease localization patterns across all datasets, with field-captured images maintaining clear attention focus despite challenging conditions, supporting practical real-world applicability.

C. Limitations and Future Work

While our method demonstrates strong performance on field datasets, it exhibits limitations in laboratory-controlled environments, achieving 24.76% accuracy on Taiwan Tomato Leaves Dataset and 23.80% on PlantVillage. This performance gap stems from the domain mismatch between our field-based training data (PlantDoc) and controlled laboratory conditions featuring standardized lighting and uniform backgrounds. Future research should explore domain adaptation techniques and multi-environment training strategies to bridge this gap. Additionally, incorporating synthetic data generation and adversarial training could enhance robustness across diverse imaging conditions. Training on balanced datasets containing both field and laboratory samples would improve applicability for comprehensive agricultural monitoring systems.

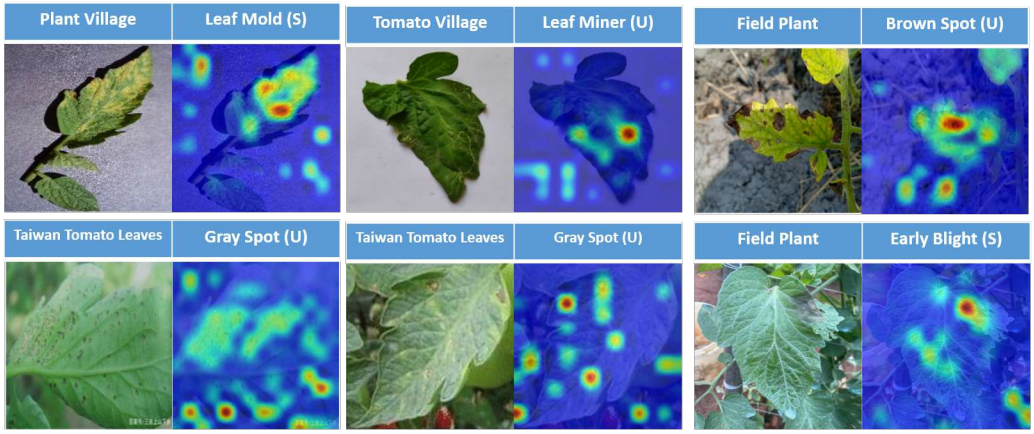

Fig. 3: Attention heatmap visualizations of our proposed model across different datasets and disease conditions. Each row shows original leaf images (left) and corresponding attention maps (right) highlighting disease-relevant regions. (S) indicates seen disease classes present in training data, while (U) denotes unseen classes not encountered during training.

VI. Conclusion

This paper presents a novel Flexible Contrastive Clustering framework for zero-shot cross-domain tomato disease classification that addresses fundamental limitations in agricultural computer vision. Our approach leverages one-to-many image-text relationships and class-based clustering to achieve superior generalization compared to traditional vision-language models.

Experimental evaluation across four diverse datasets totaling 17,313 images demonstrates the effectiveness of our method, achieving 30.15% accuracy and 28.05% weighted F1-score on average—substantially outperforming CLIP (19.99%, 15.20%), SigLIP (15.00%, 8.96%), and LongCLIP (23.13%, 18.25%). The approach excels particularly in challenging field conditions, achieving 59.70% accuracy on FieldPlant and 26.52% on Tomato Village, representing significant improvements over baseline methods.

Key contributions include: (1) the FCC framework that optimizes class-based clustering in joint embedding space, (2) dynamic training strategy utilizing multiple textual descriptions per image, (3) comprehensive zero-shot evaluation establishing new benchmarks for cross-dataset plant disease classification, and (4) interpretability through attention visualization demonstrating effective disease localization for both seen and unseen categories.

While our approach demonstrates significant improvements in cross-domain generalization, the primary limitation involves reduced performance on laboratory datasets compared to domain-specific models, indicating the need for future research into adaptive fine-tuning techniques that can maintain laboratory accuracy while preserving field robustness. Nevertheless, this work represents a substantial advancement toward practical automated plant disease detection systems capable of operating effectively across diverse real-world agricultural environments with minimal expert supervision. The combination of interpretable decision-making processes and robust cross-domain performance addresses critical requirements in precision agriculture, where expert knowledge is often limited and transparent, reliable diagnostic tools are essential for achieving widespread farmer adoption and improving global food security outcomes.

Future research directions should focus on both methodological improvements and practical agricultural applications to advance automated plant disease detection systems. From a methodological perspective, comprehensive ablation studies are needed to determine the optimal quantity of textual descriptions per disease class and their impact on model performance, while architecture optimization research should examine fine-tuning layer selection strategies to maximize transfer learning effectiveness. Additionally, investigating parameter efficiency techniques could enable achieving comparable performance with models containing fewer than 7.5 million parameters, making deployment more feasible for resource-constrained agricultural settings. On the application side, developing multi-environment training strategies that systematically incorporate diverse field conditions would enhance real-world robustness, while synthetic data augmentation approaches could address the challenge of rare disease detection where collecting sufficient training samples remains difficult. Furthermore, optimizing these models for real-time deployment and integrating them with existing precision agriculture systems would facilitate widespread adoption, ultimately bridging the gap between laboratory research and practical farm-level implementation where automated disease detection can significantly impact crop management decisions.

References

[1] FAO, *Agricultural production statistics 2010–2023*, 2024.

[2] D. Velayudhan, M. Shafay, S. G. Khawaja, N. Werghi, and T. Hassan, "Semi-supervised segmentation-driven classification pipeline for grading cassava leaf diseases," in *2024 International Conference on Engineering and Emerging Technologies (ICEET)*, 2024, pp. 1–6. DOI: 10.1109/ICEET65156.2024.10913946.

[3] S. P. Mohanty, D. P. Hughes, and M. Salathé, "Using deep learning for image-based plant disease detection," *Frontiers in Plant Science*, vol. 7, p. 1419, 2016. DOI: 10.3389/fpls.2016.01419.

[4] E. C. Too, L. Yujian, S. Njuki, and L. Yingchun, "A comparative study of fine-tuning deep learning models for plant disease identification," *Computers and Electronics in Agriculture*, vol. 161, pp. 272–279, 2019.

[5] X. Chen, G. Zhou, A. Chen, J. Yi, W. Zhang, and Y. Hu, "Identification of tomato leaf diseases based on combination of abck-bwtr and b-arnet," *Computers and Electronics in Agriculture*, vol. 178, p. 105 730, 2020.

[6] M. Shafay, T. Hassan, A. Ahmed, D. Velayudhan, J. Dias, and N. Werghi, "Programmable broad learning system to detect concealed and imbalanced baggage threats," in *2022 2nd International Conference on Digital Futures and Transformative Technologies (ICoDT2)*, IEEE, 2022, pp. 1–6.

[7] D. Velayudhan and S. Paul, "Two-phase approach for recovering images corrupted by gaussian-plus-impulse noise," in *2016 International Conference on Inventive Computation Technologies (ICICT)*, IEEE, vol. 2, 2016, pp. 1–7.

[8] M. Shafay, A. Ahmed, T. Hassan, J. Dias, and N. Werghi, "Programmable broad learning system for baggage threat recognition," *Multimedia Tools and Applications*, vol. 83, no. 6, pp. 16 179–16 196, 2024.

[9] M. Alansari, A. Ahmed, K. Alnuaimi, *et al.*, "Multi-scale hierarchical contour framework for detecting cluttered threats in baggage security," *IEEE Access*, vol. 12, pp. 77 454–77 467, 2024.

[10] E. Moupojou *et al.*, "Fieldplant: A dataset of field plant images for plant disease detection and classification with deep learning," *IEEE Access*, vol. 11, pp. 35 398–35 410, 2023. DOI: 10.1109/ACCESS.2023.3263042.

[11] D. Velayudhan, T. Hassan, E. Damiani, and N. Werghi, "Recent advances in baggage threat detection: A comprehensive and systematic survey," *ACM Computing Surveys*, 2022.

[12] A. Ahmed, M. Alansari, K. Alnuaimi, D. Velayudhan, T. Hassan, and N. Werghi, "Detection transformer framework for recognition of heavily occluded suspicious objects," in *2023 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA)*, IEEE, 2023, pp. 1–6.

[13] H. Raja, T. Hassan, M. U. Akram, and N. Werghi, "Clinically verified hybrid deep learning system for retinal ganglion cells aware grading of glaucomatous progression," *IEEE Transactions on Biomedical Engineering*, pp. 1–1, 2020. DOI: 10.1109/TBME.2020.3030085.

[14] M. Hayat, S. H. Khan, N. Werghi, and R. Goecke, "Joint registration and representation learning for unconstrained face identification," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1551–1560. DOI: 10.1109/CVPR.2017.169.

[15] A. Radford, J. W. Kim, C. Hallacy, *et al.*, "Learning transferable visual representations by contrastive language-image pre-training," in *International Conference on Machine Learning*, PMLR, 2021, pp. 8748–8763.

[16] D. Velayudhan, A. Ahmed, T. Hassan, M. Bennamoun, E. Damiani, and N. Werghi, "Context-aware transformers for weakly supervised baggage threat localization," in *2023 IEEE International Conference on Image Processing (ICIP)*, 2023, pp. 3538–3542. DOI: 10.1109/ICIP49359.2023.10221975.

[17] S. P. Mohanty, D. P. Hughes, and M. Salathé, "Using deep learning for image-based plant disease detection," *Frontiers in Plant Science*, vol. 7, p. 1419, 2016. DOI: 10.3389/fpls.2016.01419.

[18] A. Khan, T. Hassan, M. Shafay, *et al.*, "Tomato maturity recognition with convolutional transformers," *Scientific Reports*, vol. 13, no. 1, p. 22 885, 2023.

[19] M. Agarwal, A. Singh, S. Arjaria, A. Sinha, and S. Gupta, "Toled: Tomato leaf disease detection using convolution neural network," in *Procedia Computer Science*, vol. 167, Elsevier, 2020, pp. 293–301. DOI: 10.1016/j.procs.2020.03.225.

[20] A. Fuentes, S. Yoon, S. C. Kim, and D. S. Park, "A robust deep-learning-based detector for real-time tomato plant diseases and pests recognition," *Sensors*, vol. 17, no. 9, p. 2022, 2017. DOI: 10.3390/s17092022.

[21] M. Arsenovic, M. Karanovic, S. Sladojevic, A. Anderla, and D. Stefanovic, "Solving current limitations of deep learning based approaches for plant disease detection," *Symmetry*, vol. 11, no. 7, p. 939, 2019. DOI: 10.3390/sym11070939.

[22] J. G. A. Barbedo, "Factors influencing the use of deep learning for plant disease recognition," *Biosystems Engineering*, vol. 172, pp. 84–91, 2018. DOI: 10.1016/j.biosystemseng.2018.05.013.

[23] W. B. Demilie, "Plant disease detection and classification techniques: A comparative study of the performances," *Journal of Big Data*, vol. 11, no. 1, pp. 1–29, 2024. DOI: 10.1186/s40537-023-00863-9.

[24] Z. Wang, Z. Wu, D. Agarwal, and J. Sun, "Medclip: Contrastive learning from unpaired medical images and text," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, vol. 2022, 2022, p. 3876.

[25] M. Li, M. Meng, M. Fulham, D. D. Feng, L. Bi, and J. Kim, "Enhancing medical vision-language contrastive learning via inter-matching relation modelling," *IEEE Transactions on Medical Imaging*, 2025.

[26] X. Wu, X. Fan, P. Luo, S. D. Choudhury, T. Tjahjadi, and C. Hu, "From laboratory to field: Unsupervised domain adaptation for plant disease recognition in the wild," *Plant Phenomics*, vol. 5, p. 0038, 2023. DOI: 10.34133/plantphenomics.0038.

[27] C. Hu, X. Wu, X. Fan, P. Luo, and T. Tjahjadi, "Multi-representation subdomain adaptation network with uncertainty regularization for cross-species plant disease classification," *Plant Phenomics*, vol. 5, p. 0038, 2023. DOI: 10.34133/plantphenomics.0038.

[28] K. Zhan, Y. Peng, M. Liao, and Y. Wang, "Domain generalization plant leaf disease recognition: Toward from laboratory to field," *Engineering Applications of Artificial Intelligence*, vol. 138, p. 109 698, 2024. DOI: 10.1016/j.engappai.2024.109698.

[29] A. Dolatabadian, P. E. Bayer, S. Tirnaz, B. Hurgobin, D. Edwards, and J. Batley, "Image-based crop disease detection using machine learning," *Plant Pathology*, 2025. DOI: 10.1111/ppa.14006.

[30] M. Shoaib, B. Shah, S. Ei-Sappagh, *et al.*, "An advanced deep learning models-based plant disease detection: A review of recent research," *Frontiers in Plant Science*, vol. 14, p. 1 158 933, 2023. DOI: 10.3389/fpls.2023.1158933.

[31] D. Singh, N. Jain, P. Jain, P. Kayal, S. Kumawat, and N. Batra, *Plantdoc: A dataset for visual plant disease detection*, 2020. arXiv: 1912.10317 [cs.CV].

[32] M.-L. Huang and Y.-H. Chang, *Dataset of tomato leaves*, version V1, 2020. DOI: 10.17632/ngdgg79rzb.1.

[33] M. Gehlot, R. Saxena, and G. Gandhi, "Tomato-village: A dataset for end-to-end tomato disease detection in a real-world environment," *Multimedia Systems*, vol. 29, pp. 3305–3328, 2023. DOI: 10.1007/s00530-023-01158-y.

[34] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," *arXiv preprint arXiv:2303.15343*, 2023, SigLIP model reference.

[35] B. Zhang, Q. Gu, Y. Wu, *et al.*, "Longclip: Scaling clip with long text inputs," *arXiv preprint arXiv:2403.15378*, 2023, LongCLIP model reference.