

Article

Unseen-Crop Plant Disease Classification via Disentangled Representation Learning

Zhenzhen Wu ¹, Jianli Guo ¹, Wei Hou ², Kun Zhou ¹, Kerang Cao ³ and Hoekyung Jung ^{4,*}

¹ New Generation Information Technology Research Center, Weifang University of Science and Technology, Weifang 262700, China; wuzhenzhen@wfust.edu.cn (Z.W.); guojianli@wfust.edu.cn (J.G.); zhokun2025@126.com (K.Z.)

² The Youth Innovation Team of Shaanxi Universities, Shaanxi Polytechnic University, Hanzhong 723000, China; 13572231943@126.com

³ Key Laboratory of Intelligent Technology of Chemical Process Industry in Liaoning Province, College of Computer Science and Technology, Shenyang University of Chemical Technology, Shenyang 110142, China; caokerang@syuct.edu.cn

⁴ Computer Engineering Department, Paichai University, Daejeon 35345, Republic of Korea

* Correspondence: hkjung@pcu.ac.kr

Abstract

Deep learning has accelerated progress in plant disease recognition, providing strong technical support for early diagnosis and precision management. However, models often lack robustness and generalization when confronted with novel crops absent from the training set, leading to a marked performance drop in cross-unseen-crop scenarios. Cross-crop generalization for plant disease recognition requires models to identify known disease categories in crop domains never observed during training. A central challenge is that disease symptoms are strongly coupled with crop-specific appearance cues, which severely degrades generalization. Here, TDC (Text-guided feature Disentanglement Contrast) is introduced as a feature-disentanglement framework for cross-crop plant disease recognition. The proposed method employs a dual-branch visual encoder to separately capture disease semantic representations and crop-domain representations, and it leverages a frozen CLIP text encoder to use disease and crop prompts for text-guided semantic anchoring. A semantic-anchor-only contrastive disentanglement strategy is further formulated under a hybrid label space, where crop-branch features are incorporated as stop-gradient hard negatives to suppress semantic–domain information leakage and strengthen the intra-class aggregation of the same disease across crops. Residual domain-discriminative cues are mitigated via domain-adversarial learning. During inference, only the disease branch is retained for classification, improving generalization while reducing deployment overhead. Experiments demonstrate that under the PlantVillage cross-crop setting, the method achieves 98.04% and 74.29% Top-1 accuracy on seen and unseen crop domains, respectively. Moreover, it attains 81.99% on a real-world field dataset of strawberry powdery mildew and 76.31% on a low-illumination degradation set, validating robustness under realistic imaging distribution shifts.



Academic Editor: George A. Tsihrintzis

Received: 5 March 2026

Revised: 1 April 2026

Accepted: 5 April 2026

Published: 8 April 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: cross-crop generalization; disentangled representation learning; vision–language alignment; contrastive learning; plant disease classification

1. Introduction

With continued global population growth and accelerating climate change, ensuring food security has become an urgent worldwide priority. Plant diseases are among the

primary threats to crop yield and quality, imposing substantial losses on agricultural productivity and economic returns [1,2]. In practice, disease diagnosis still largely relies on visual assessments by experts—an approach that is labor-intensive, time consuming, and increasingly constrained by a shortage of trained personnel [3]. Improving the accuracy and efficiency of disease classification has therefore become a pressing objective in modern agricultural research, motivating intensive efforts to build accurate and efficient automated disease recognition systems at the intersection of agriculture and computer vision [4]. Although leaf symptoms are not the earliest manifestation for all plant diseases, they constitute one of the first observable signs for a large proportion of diseases. Consequently, leaf imagery provides an important basis for the early detection of many plant diseases [5]. Early detection and classification from leaf images are critical for enhancing agricultural productivity and strengthening risk early-warning capabilities [6].

In recent years, deep learning has advanced rapidly in plant disease detection and recognition [7,8]. Karthikeyan et al. [9] developed a DenseNet-based framework using deep transfer learning for the content-based retrieval of agricultural plant disease images, aiming to improve recognition accuracy and retrieval efficiency. To address the diversity issue in plant disease classification, Eunice et al. [10] developed a cross-crop disease detection method that integrates multi-domain data augmentation and transfer learning with the aim of improving the learning of disease representations across different crop types. Based on multi-crop leaf images collected under relatively controlled conditions, their model achieved good classification results, indicating that the method can effectively capture cross-crop differences in leaf disease data.

Despite recent progress, most plant disease classification methods assume that training and test data share the same crop domains and disease label space. Here, we instead consider a more challenging and practically relevant setting in which disease categories are seen during training, whereas the crop domain at test time is unseen. The goal is therefore not to recognize novel diseases but rather to generalize disease recognition across crop species. Such a setting better reflects real agricultural scenarios, where the exhaustive annotation of all crop–disease combinations is unrealistic. Moreover, the same disease category may not be observed for every crop species during training, leading to missing crop–disease combinations in the training set. Consequently, at test time, the model may need to recognize a known disease category in a crop domain that has never been seen during training. This poses a fundamental challenge for cross-crop generalization.

The central difficulty in cross-crop generalization lies in a pronounced domain shift and the strong coupling between disease semantics and crop-appearance cues. Because crop morphology, texture, and imaging background are often correlated with disease labels during data collection, models can easily learn shortcut features tied to crop appearance or acquisition conditions. When transferring from source crop domains to a new crop domain, such domain-dependent representations often fail, leading to a substantial performance drop in cross-crop testing. This highlights the limited adaptability of coupled crop–disease features in disease recognition. We promote the aggregation of shared disease labels across crops and the separation of distinct disease categories in a disease semantic space. Crop-specific variation is constrained to an independent crop-domain subspace, as shown in Figure 1.

Motivated by these challenges, this paper focuses on plant disease recognition under unseen-crop generalization. During training, the model observes only a set of seen crop domains, whereas at test time, it must recognize seen disease concepts in unseen crop domains. This setting closely matches practical needs: a disease may migrate to, or be newly recorded on, a previously unobserved crop species, while collecting sufficient labeled samples for every crop–disease combination is unrealistic in both cost and time. Consequently, learning

feature representations with strong generalization ability has become an important research direction in this field.

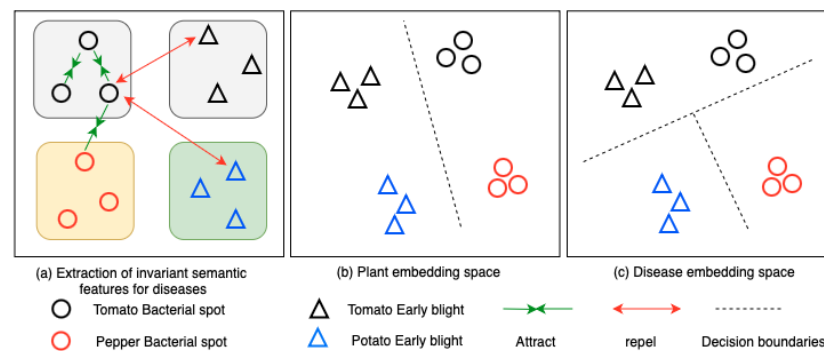


Figure 1. (a) Extracting disease-related, domain-invariant semantic representations, where samples of the same disease are closer and different diseases are more separable. (b) Correctly classifying disease categories in the disease-embedding space. (c) Correctly classifying crop domains in the plant-embedding space.

We note that domain generalization and disentangled representation learning provide a principled framework for mitigating crop-domain shift by learning domain-invariant semantic representations from multiple source domains. Domain generalization methods seek to improve robustness to domain variation through feature alignment and regularization, invariant risk minimization, meta-learning strategies, and style perturbation [11]. Disentangled representation learning [12] explicitly separates the underlying generative factors within samples, thereby reducing the interference of domain-specific nuisance factors in category discrimination and improving generalization to unseen target domains. However, existing methods still show clear limitations in cross-crop plant disease recognition. Many methods implicitly assume similar crop distributions in the training and test sets. Some methods require additional annotations. Others are limited to coarse binary recognition. In this task, disease symptoms are closely coupled with crop-specific appearance cues. Domain-invariant learning in the visual space does not explicitly remove semantic–domain coupling. It may even weaken visual cues that are critical for disease discrimination. As a result, crop-domain information can still be encoded into the learned disease representation.

To address these challenges, TDC (Text-guided feature Disentanglement Contrast) is introduced as a text-guided framework for contrastive feature disentanglement. Cross-crop disease recognition is formulated from the perspective of domain generalization (DG): during training, only a set of seen crop domains is observable, whereas at test time, seen disease categories must be recognized in unseen crop domains. TDC employs dual-branch visual encoders to separately extract disease semantic representations and crop-domain representations. To construct a transferable disease semantic space, a frozen CLIP text encoder provides disease and crop textual prompts as semantic anchors, enabling cross-modal alignment and concept-level constraints. To suppress semantic–domain coupling and prevent information leakage, a semantic-anchor-only contrastive disentanglement objective in a mixed label space is further developed in which crop-branch features are incorporated as stop-gradient hard negatives. This design discourages disease representations from absorbing crop-domain information at the optimization level. Domain-adversarial learning is employed to eliminate residual crop-domain-discriminative information. During inference, only the disease branch is retained for classification, thereby balancing generalization capability with deployment efficiency. The framework addresses a setting with shared disease labels and unseen crop domains. Text-guided disease prompts are used as semantic anchors for cross-crop alignment. Crop-domain variation is explicitly separated from

disease semantics. A semantic-anchor-only contrastive disentanglement is further designed for the mixed crop-disease label space. This formulation targets semantic-domain coupling. The main contributions are summarized as follows:

- The cross-crop plant disease recognition task is systematically defined from a domain generalization perspective. Under the setting where crop domains are unseen but disease categories are seen, disease-specific domain-invariant representations are learned, substantially reducing the reliance on crop-appearance cues and enabling disease-semantic-centered disease classification across plant species, thereby markedly improving generalization in multi-plant disease classification.
- A text-guided semantic anchoring module is introduced. Disease and crop textual prompts constrain the disease and crop branches, respectively, mapping visual disease features into a shared, semantically invariant concept space. This promotes the structured separation of semantic and domain factors and improves cross-domain transferability and interpretability.
- A semantic-anchor-only contrastive disentanglement strategy is proposed. Within a mixed label space, same-disease aggregation across crops and different-disease separation are strengthened, while semantic–domain coupling and information leakage are explicitly weakened by treating crop-branch representations as stop-gradient hard negatives, thereby enhancing robust recognition in unseen crop domains.

In summary, TDC builds on a dual-branch architecture and combines text-guided semantic anchoring with mixed-label-space contrastive constraints to enable the stable extraction of disease semantics in cross-crop scenarios. By retaining only the disease branch at the inference time, deployment overhead is reduced while strong generalization performance is preserved. The remainder of this paper is organized as follows. Section 2 reviews related work; Section 3 presents the problem formulation, data construction, and the network architecture, loss functions, and training procedure of TDC; Section 4 reports comparative experiments and ablation studies with visualization-based analyses; Section 5 discusses methodological strengths, limitations, and potential applications; and Section 6 concludes the paper and outlines directions for future work.

2. Related Work

2.1. Disease Recognition in Unseen Crop Domains

Deep learning has been extensively adopted for plant disease recognition. Recent surveys have summarized advances in plant disease detection based on state-of-the-art deep learning models with emphasis on the performance of different network architectures and feature-learning strategies across diverse recognition scenarios, commonly used datasets and evaluation protocols, key challenges, and future research directions [13]. Recent studies have begun to address the distribution shift in plant disease recognition more explicitly. Wu et al. [14] proposed an unsupervised domain adaptation framework for plant disease recognition from laboratory to field settings. Their method improved cross-domain recognition by combining multi-representation learning, subdomain adaptation, and uncertainty regularization. Gao et al. [15] further investigated multi-source domain adaptation for potato disease recognition in field environments. Their model addressed the domain mismatch caused by illumination variation and heterogeneous acquisition conditions. Yang et al. [16] examined cross-domain few-shot learning for crop disease identification and showed that performance drops markedly when source and target domains differ. More recently, Zhan et al. [17] introduced domain generalization benchmarks for plant leaf disease recognition and evaluated generalization to unseen domains. These studies confirm the importance of cross-domain robustness in plant disease analysis. However, most existing methods

still focus on domain adaptation, few-shot transfer, or benchmark-specific settings. Robust recognition under unseen crop domains therefore remains insufficiently explored.

To meet the demand for cross-crop generalization, several studies have attempted to weaken reliance on crop–disease combination labels. Bouacida et al. [18] partitioned leaf images into small patches, labeled each patch as either diseased or healthy, and classified the patches using a lightweight Inception network. This design avoids explicit distinctions among crop species or disease types, alleviating the dependence of deep models on specific crop–disease pairs and enabling efficient cross-crop, cross-disease detection and severity assessment. Nevertheless, the formulation is primarily a binary discrimination (healthy vs. unhealthy) and is therefore insufficient for fine-grained identification across multiple crops and multiple diseases. Another line of work links leaf-image features with expert-defined binary semantic attributes, constructing a zero-shot classification framework capable of recognizing disease categories unseen during training [19]. Although this framework can support cross-species disease classification, its performance is contingent on the choice and quality of the semantic attributes, and accuracy may decline when visual differences among crops are pronounced. In addition, per-image attribute annotation incurs substantial extra cost.

Crop–disease recognition has also been formulated as Compositional Zero-Shot Learning (CZSL) [20,21]. This setting focuses on unseen crop–disease compositions while assuming that both crop concepts and disease concepts are observed during training [22,23]. In contrast, the present task targets seen diseases in unseen crop domains. Its core difficulty lies in domain shift, not merely missing concept pairs. As a result, CZSL does not fully address the cross-crop generalization problem considered here.

2.2. Disentangled Representation Learning for Cross-Domain Robustness

Domain generalization has long focused on maintaining stable model performance under domain shift. In this setting, models are trained on data from multiple source domains and are expected to generalize to unseen target domains. Disentangled representation learning aims to separate different generative factors of the data into distinct subspaces, thereby reducing the influence of nuisance factors on task discrimination. Related surveys have noted that disentanglement can be achieved through mechanisms such as reconstruction-based constraints, adversarial learning, or information bottlenecks, and it has demonstrated value in tasks including cross-domain recognition. Peng et al. [24] proposed a domain-irrelevant learning framework that mitigates interference from domain factors via disentangled representations. The deep adversarial disentangled autoencoder DADA (Deep Adversarial Disentangled Autoencoder) further promotes the separation of domain-irrelevant factors through an adversarial mechanism in a unified procedure. Lin et al. [25] introduced an enhanced feature disentanglement module; by combining disentanglement with response alignment, the approach effectively alleviates performance bottlenecks caused by feature mismatch in both classification and regression tasks. Zhang et al. [26] proposed DisMAE, integrating masked autoencoders with disentanglement learning for unsupervised domain generalization and achieving competitive performance on cross-domain vision tasks. These methods improve domain robustness, but they are not sufficient for the current task. They are mostly developed for general cross-domain vision settings. They do not explicitly address the semantic coupling between disease symptoms and crop-domain cues. They also do not target the recognition of seen diseases in unseen crop domains. This leaves the cross-crop disease recognition problem only partially solved.

Nevertheless, disentanglement learning in vision often suffers from an under-constrained learning objective. With only visual reconstruction or adversarial constraints, models may still encode mixtures of class-relevant and domain-relevant information, result-

ing in subspaces with limited semantic purity and, consequently, constrained interpretability and transferability. This issue is particularly acute for cross-crop disease recognition. In settings where crop-appearance cues are strongly coupled with disease symptoms, pursuing domain invariance solely within the visual feature space makes it difficult to explicitly separate disease semantics from crop-appearance variation, thereby undermining the stability of disease-centric representations. Moreover, overly strong domain-invariance constraints may suppress not only domain-specific noise but also critical disease-discriminative cues, leading to suboptimal generalization on unseen crop domains.

2.3. Vision–Language Multimodality and Textual Semantic Anchors

To mitigate semantic–domain information leakage during disentanglement and to enhance the purity and cross-domain consistency of the semantic subspace, it is essential to introduce cross-domain-stable semantic anchors that impose targeted constraints on the representation space. Semantic priors provided by vision–language pretrained models can more effectively regularize the structure of representations learned during disentanglement. CLIP (Contrastive Language–Image Pretraining) [27] performs contrastive learning on large-scale image–text pairs to jointly train image and text encoders, enabling the use of natural-language supervision to produce highly transferable visual representations and yielding substantial gains in zero-shot and few-shot classification. In addition, Huang et al. [28] improved multimodal domain generalization by disentangling representations into modality-shared semantics and modality-specific domain information. Jiang et al. [29] fuse CLIP’s global semantic representations with DINOv2’s multi-scale structural features via cross-modal integration to enable zero-shot anomaly detection in industrial scenarios. Zi et al. [30] recast unsupervised dense counting as an image–text matching problem and leverage multimodal deep shared prompt tuning together with cross-modal alignment ranking to exploit the complementarity between CLIP’s visual and textual modalities, thereby improving unsupervised counting performance. Together, these studies indicate that incorporating language-driven semantic constraints can help structure representation spaces and support generalization.

In agricultural vision, multimodal learning has also attracted attention, particularly for complex pattern classification and automated feature extraction. Phan et al. [31] presented a multifaceted vision–language pretraining framework that enhances pathological detection by decomposing disease descriptions. Nevertheless, such approaches often still depend on extensive annotations or powerful pretrained models, which may be constrained by computational resources in practical applications. While multimodal fusion is advantageous, cross-species classification remains challenging. Liaw et al. [32] applied text-guided mechanisms to crop disease recognition. By leveraging language priors to improve distributional consistency in a synthesized feature space, this approach effectively improved recognition performance for unseen crop–disease combinations. Overall, these studies suggest that combining CLIP’s multimodal alignment capability with text-guided mechanisms can bridge visual and semantic gaps, offering strong support for cross-domain disease classification.

Inspired by these lines of work, our objective is to alleviate the performance degradation caused by distribution shifts in cross-crop disease recognition by explicitly disentangling disease semantic representations from domain-variant factors. This strategy aims to guide models to prioritize disease-discriminative cues that remain stable across crop domains while suppressing crop-domain-related interference in feature learning and decision boundaries, thereby improving generalization and robustness on unseen crop domains and in real-world application settings.

Following the above review, this paper focuses on cross-crop plant disease recognition in unseen crop domains. The proposed framework introduces text-guided semantic anchoring to improve cross-domain semantic consistency. It also incorporates a contrastive disentanglement design to separate disease-related cues from crop-domain variation. As a result, the method reduces semantic–domain coupling and improves recognition robustness under domain shift.

3. Materials and Methods

3.1. Problem Formulation

Cross-crop disease recognition aims to learn disease-discriminative representations that are transferable across species, enabling the knowledge of disease categories observed during training to generalize to crop species absent from training. The objective is to accurately recognize known diseases in unseen crop domains, thereby achieving generalizable disease recognition for new species.

Under the cross-domain generalization setting for recognizing seen disease categories in unseen crop domains, let C denote the total number of disease categories and M the number of crop domains. Without loss of generality, the unseen target domain is indexed as the M -th domain, while the remaining $M - 1$ domains are treated as seen domains. For the j -th crop domain, the sample set is defined as $\mathcal{X}_j = \{(x_i^j, y_i^j)\}_{i=1}^{n_j}$, where n_j denotes the number of samples in domain j , x_i^j denotes an input image, and $y_i^j \in \{1, \dots, C\}$ denotes the corresponding disease label. The unseen-domain test set is defined as $\mathcal{X}_{\text{test}}^{\text{unseen}} = \mathcal{X}_M = \{(x_i^M, y_i^M)\}_{i=1}^{n_M}$, while the seen-domain test set is defined as $\mathcal{X}_{\text{test}}^{\text{seen}} = \bigcup_{j=1}^{M-1} \mathcal{X}_j^{\text{seen}}$, where $\mathcal{X}_j^{\text{seen}}$ denotes the test subset of the j -th seen crop domain. The model is trained on seen crop domains and evaluated on both seen-domain and unseen-domain test sets with all disease labels remaining in the same category space. The goal is to learn disease-discriminative representations that remain stable under crop-domain shift and support reliable recognition in unseen crop domains.

To address disease recognition in dynamic agricultural scenarios with unseen crop domains, a text-guided feature disentanglement contrast framework, termed TDC, is proposed, as shown in Figure 2. The framework is designed to mitigate the strong coupling between disease semantics and crop-domain variation. It learns disease-discriminative representations while suppressing crop-related nuisance factors. Text-guided semantic anchoring provide stable cross-domain guidance, and disentanglement learning further reduces semantic–domain leakage. This design improves the transferability of disease representations across crop domains and supports reliable disease classification in unseen crop domains.

3.2. Segment–Raw Cross-View Contrastive Enhancement and Background Randomization

To obtain domain-invariant disease semantic representations under cross-crop settings, a two-view augmentation strategy is designed to address background bias in the PlantVillage dataset, where raw images exhibit highly homogeneous backgrounds. This strategy strengthens robustness to crop-domain-related perturbations. Because raw images share strongly homogeneous backgrounds, contrastive alignment can encourage reliance on shortcut cues such as background consistency or boundary shape, thereby aggravating semantic–domain entanglement. To counteract the background bias caused by such homogeneity, a structured cross-view learning mechanism is adopted. Specifically, a segmentation-based view is introduced and combined with background randomization to explicitly suppress background-induced spurious correlations. For view construction, two-view augmentation is used to improve coverage and robustness in contrastive learning. Each raw image yields two views that jointly participate in contrastive learning. The

Segment View serves as a structured alternative view of the same instance and is paired with an augmented raw view as two observations of that instance. The augmented raw view $\hat{x}^{\text{raw}}$ retains the original background information and provides rich contextual textures. The background-randomized segmentation view $\hat{x}^{\text{seg}}$ is a structured alternative observation of the same instance. To address the highly homogeneous background in PlantVillage raw images, background randomization is applied to the region outside the segmented foreground. Under this construction, $\hat{x}^{\text{raw}}$ and $\hat{x}^{\text{seg}}$ share the same disease-instance foreground semantics, while their background distributions are fully independent and may even be mutually exclusive. This design forces the model, during cross-view alignment, to discard background noise and focus on the shared disease semantic cues, thereby explicitly suppressing background-induced spurious correlations and ensuring that cross-view consistency is primarily driven by lesion semantics. On the disease semantic branch, a class-supervised contrastive objective is adopted to learn disease representations that exhibit intra-class compactness and inter-class separability. Under this objective, $\hat{x}^{\text{seg}}$ and $\hat{x}^{\text{raw}}$ are not treated as the only positive pair. Instead, they are regarded as two sample views from the same disease category and are included in the sample set, forming an intra-class alignment constraint jointly with disease representations of all same-class samples. For each disease semantic representation obtained from an augmented view, all representations in the batch are denoted as D_i with label y_i . The set of positives for D_i consists of all representations satisfying $y_k = y_i$, while representations from other categories are treated as negatives. Defining positives via disease-category consistency enables a more stable learning of cross-crop-consistent disease semantics and helps suppress background bias. The resulting two-view representations are optimized under the mixed-label-space cross-branch contrastive loss $\mathcal{L}_{\text{msc}}$, whose formal definition is given in Equation (3). In this formulation, the two views are not treated as an isolated positive pair. Instead, they are incorporated into the class-level positive set and jointly contribute to intra-class compactness and inter-class separation. During training, $\mathcal{L}_{\text{msc}}$ is further combined with the cross-modal alignment losses and the domain-adversarial loss in the overall objective, as defined in Equation (5). Augmentations applied to raw images include cropping, color jittering, grayscale conversion, random Gaussian blur, and random erasing. All views are normalized. In parallel, the corresponding Segment View is constructed for each raw view and subjected to background randomization. During training, each sample is represented by two synchronized inputs, namely one raw view and one segmentation view derived from the same image. This design is intended to suppress shortcut learning caused by the highly homogeneous background in PlantVillage. The raw view is augmented with random resized cropping (scale = 0.7–1.0), color jittering (brightness/contrast/saturation/hue = 0.4/0.4/0.4/0.1), random grayscale conversion ($p = 0.2$), and Gaussian blur ($p = 0.3$, sigma = 0.1–2.0). Normalization is applied at the end of the pipeline. The segmentation view uses the same geometric augmentation. Outside the leaf foreground, full background randomization is performed. The randomized background is sampled from three sources, namely solid-color backgrounds, Gaussian-noise backgrounds, and blurred natural backgrounds, with a ratio of 0.4:0.3:0.3. A boundary feathering operation with a width of 3 pixels is further applied. This setting suppresses background-driven shortcut learning in PlantVillage while preserving lesion-foreground semantics. All operations were implemented using PyTorch (v2.6.0) in Python (v3.13.2).

3.3. Multimodal Feature Extraction

A dual-branch feature extraction architecture was adopted to disentangle disease-related and crop-related representations. This design reduces the coupling of the two factors in a shared feature space. For the visual modality, the disease branch employs a ViT-

B/16 backbone with 12 Transformer layers, whereas the crop branch employs a ViT-B/16 backbone with 6 Transformer layers [33]. Both branches were initialized from the pretrained CLIP visual encoder. For the textual modality, a frozen CLIP text encoder was used to provide stable semantic priors. This strategy helps preserve the cross-modal knowledge acquired during large-scale image-text pretraining. It is particularly beneficial under limited training data. It also improves generalization across different crop categories.

The pretrained parameters were obtained from the official OpenAI CLIP implementation (GitHub repository: openai/CLIP, commit dcba3cb; OpenAI, San Francisco, CA, USA).

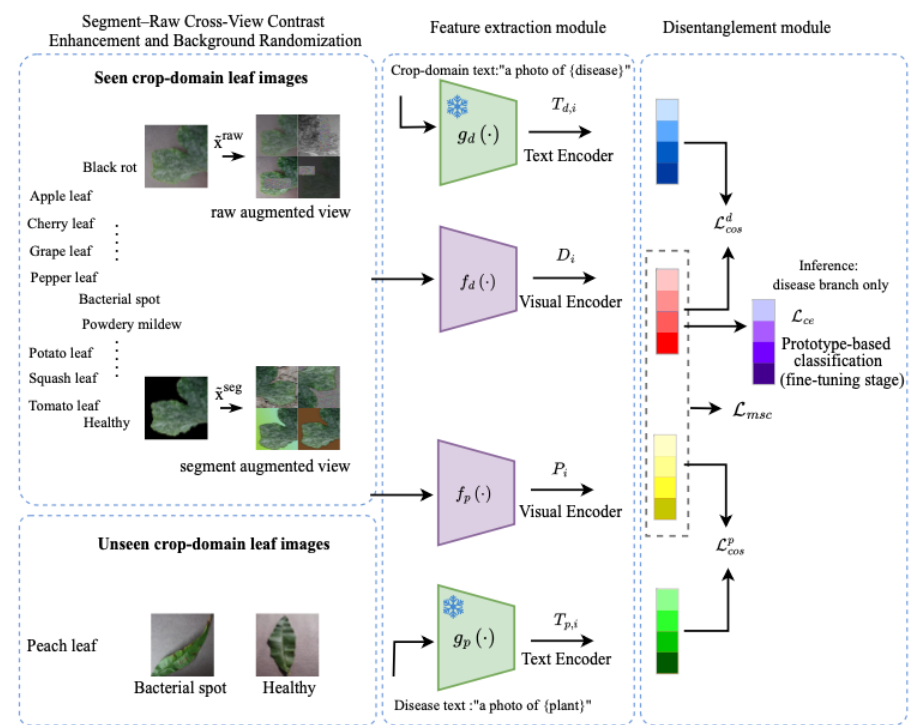


Figure 2. Overview of the proposed framework. During training, each sample includes two synchronized inputs from the PlantVillage subset: an original leaf image with background and its corresponding segmented leaf image. The original image is used as the raw view with visual augmentation. The segmented image is used as the segment view with background randomization. These two inputs are fed into the disease branch and the crop branch to learn disease-semantic and crop-domain representations. On the text side, a frozen CLIP text encoder maps disease prompts and crop prompts into semantic anchors for cross-modal alignment. In the mixed label space, the disease representation serves as the only trainable anchor, while the crop representation is used as a stop-gradient hard negative. Domain-adversarial learning is further applied to reduce residual crop-related factors. During inference, only the disease branch is retained, and disease concepts are predicted using class prototypes. The crop labels denote crop domains, and the visual examples correspond to leaf images. Different colors are used to distinguish the visual encoder and the text encoder for clarity. The leaf images shown are representative examples only.

For each training image x_i , the two visual branches are applied in parallel to obtain a disease visual feature D_i and a crop visual feature P_i . Meanwhile, the corresponding disease description and crop description are fed into frozen text feature extractors $g_d(\cdot)$ and $g_p(\cdot)$, respectively, yielding the text embeddings $T_{d,i}$ and $T_{p,i}$. These cross-modal embeddings provide comparable semantic anchors for subsequent alignment, enabling a more robust separation of disease semantics from crop-domain information. Under contrastive constraints, alignment can be performed more precisely, thereby supporting the robust generalization of disentangled representations to unseen species. A dual-encoder configuration is considered preferable to a single-encoder alternative. Dual-branch disentangled

modeling is more conducive to learning disease representations that are both crop-domain invariant and discriminative. In contrast, collapsing the architecture into a single encoder would entangle crop-domain information and disease-category information within the same representation space, making it difficult to distinguish domain-invariant factors that benefit classification from domain-related but category-irrelevant nuisance factors, thereby weakening cross-crop generalization. By modeling crop-domain-specific factors and disease semantics separately, the dual-branch framework suppresses interference from crop-domain bias on disease discrimination and, in turn, promotes the learning of transferable disease features. Therefore, the dual-branch design constitutes a key component for improving cross-domain classification performance in this task.

3.4. Disentangled Representation Learning

3.4.1. CLIP Text-Guided Feature Disentanglement

Because the same disease type can exhibit substantial differences in texture, shape, and color across leaves from different crops, extracting domain-invariant disease semantic representations using visual cues alone remains challenging. In recent years, large-scale vision–language models have provided a new perspective for mitigating this difficulty. Purely visual models often emphasize pixel-level details or local structures and thus are more prone to learning shallow features correlated with background, illumination, or imaging viewpoint while being less effective at capturing high-level task-relevant semantics. In contrast, introducing textual descriptions can provide higher-level semantic priors, guiding models to focus on semantic-invariant factors tightly related to the task rather than being limited to low-level cues such as shape and texture. With the development of a pretrained language model, text embedding spaces contain rich and domain-agnostic semantic knowledge that can serve as semantic references for visual learning [34]. Aligning the feature distributions of visual branches to the text embedding space can compensate for semantic information that may be overlooked during purely visual training. Moreover, textual descriptions typically emphasize the core attributes of target objects and naturally downplay nuisance factors such as illumination, background, and viewpoint, thereby suppressing semantically irrelevant noise during training and yielding more discriminative representations. Motivated by these observations, a text-guided semantic anchoring strategy is adopted. Visual and textual feature branches operate in parallel, using disease text descriptions as semantic anchors to continuously constrain and guide the learning of visual representations from the early stages of training. This strategy facilitates the stable learning of disease semantics that remain consistent across crop domains. By allowing disease semantics and crop semantics to jointly regulate visual representations, features are mapped into a more general and more discriminative disease–crop semantic space. Contrastive losses are employed to promote the disentanglement and separation of visual features, thereby improving robustness and generalization for disease classification in unknown crop domains. Specifically, a frozen CLIP text encoder is used to map predefined prompt templates, “a photo of a {Disease}” and “a photo of a {Plant}”, into semantic anchors for cross-modal alignment. To construct textual semantic anchors, simple category-name prompts are adopted. This design is intentionally lightweight. It preserves the core semantic label while avoiding excessive lexical bias from manually expanded descriptions. Short prompts are also more consistent with the label granularity used in classification. This helps maintain a stable alignment target for the disease and crop branches. In addition, compact templates reduce the risk that handcrafted attribute words dominate the text embedding and introduce irrelevant prior cues. A comparative evaluation of different prompt designs is provided in Section 4.3, where the category–name template achieves the best unseen-domain accuracy. For the i -th sample, the disease and crop prompts are fed

into the frozen text encoders $g_d(\cdot)$ and $g_p(\cdot)$ to obtain the corresponding textual semantic embeddings $T_{d,i}$ and $T_{p,i}$. Meanwhile, the image is processed by two visual branches to extract the disease feature D_i and the crop feature P_i . A cosine-similarity-based loss is adopted to enforce cross-modal alignment, encouraging the visual features to approach their semantic anchors. The loss between the disease feature D_i and its textual embedding $T_{d,i}$ is defined as

$$\mathcal{L}_{cos}^d = \sum_{i=1}^N \left(1 - \frac{T_{d,i} \cdot D_i}{\|T_{d,i}\| \|D_i\|} \right), \quad (1)$$

and the loss between the crop feature P_i and its textual embedding $T_{p,i}$ is defined as

$$\mathcal{L}_{cos}^p = \sum_{i=1}^N \left(1 - \frac{T_{p,i} \cdot P_i}{\|T_{p,i}\| \|P_i\|} \right). \quad (2)$$

where D_i and P_i denote the disease and crop visual feature vectors for the i -th sample, $T_{d,i}$ and $T_{p,i}$ are the corresponding textual semantic anchors, $\cdot$ denotes the inner product, and $\|\cdot\|$ denotes the ℓ_2 norm. Under these constraints, visual features are continuously pulled toward their corresponding positions in the semantic space during training, enabling stable cross-modal semantic alignment.

3.4.2. Semantic-Anchor-Only Contrastive Disentanglement Strategy

CLIP-style methods remain challenged in cross-crop disease recognition. Although textual prompts provide semantic guidance, without explicit constraints, it is difficult to preserve the purity of the disease semantic subspace, and semantic–domain information leakage may still occur. Therefore, merely incorporating CLIP is insufficient to guarantee robust cross-crop generalization; it must be jointly designed with structured disentanglement and objective construction. In addition, agricultural disease datasets are constrained by acquisition cost and scene complexity, and real-world data distributions are often markedly imbalanced. Limited and biased data can exacerbate the coupling between disease semantics and crop-domain factors, biasing the discrimination process toward domain-specific cues, inducing domain dependence, and weakening generalization. To address these issues, a mixed-label-space semantic-anchor-only contrastive disentanglement strategy is introduced, which is inspired by the mixed feature space construction in Chen et al. [35]. A mixed label space refers to incorporating both disease-branch features D_i and crop-branch features P_i into the same contrastive set for similarity computation, thereby explicitly characterizing the similarity structure and separability between disease semantics and crop-domain factors. Unlike conventional contrastive learning that constructs positive and negative samples only within a single representation space, crop-domain features are further included as hard negatives in similarity computation to strengthen the disease representation's repulsion to crop cues. Gradients are blocked via $\text{sg}(\cdot)$ such that the optimization signal from this contrastive objective is back-propagated only to the disease branch. Consequently, the disease branch is forced to learn stable discriminative cues that are distant from crop-domain information and to more tightly aggregate same-disease samples across crops, thereby suppressing semantic–domain information leakage at the mechanism level. Since inference targets only disease-category discrimination, the crop branch mainly serves to represent domain factors and facilitate disentanglement during training. Accordingly, retaining only the disease branch at inference suffices for classification and can substantially reduce the parameter count and computational overhead. Using disease semantic anchors obtained from the text encoder as a unified reference, semantic alignment and intra-class compactness constraints are imposed on the disease subspace

within a contrastive learning framework. Specifically, the disease-invariant semantic feature $f_d(i)$ is used as the anchor, and a contrastive objective is constructed jointly with the corresponding disease features and crop features in the same mini-batch. The crop feature $f_p(i)$ participates in similarity computation as a hard negative but does not receive gradients. A mixed-label-space cross-branch contrastive loss $\mathcal{L}_{msc}$ is defined as

$$\mathcal{L}_{msc} = \sum_{i \in \mathcal{I}} \frac{-1}{|\mathcal{P}(i)|} \sum_{j \in \mathcal{P}(i)} \log \frac{\exp(D_i \cdot \text{sg}(D_j)/\tau)}{\exp(D_i \cdot \text{sg}(D_j)/\tau) + \sum_{k \in \mathcal{N}_d(i)} \exp(D_i \cdot \text{sg}(D_k)/\tau) + \sum_{a \in \mathcal{I}} \exp(D_i \cdot \text{sg}(P_a)/\tau)} \quad (3)$$

where $\cdot$ denotes the inner product, τ is a scalar temperature parameter, $\text{sg}(\cdot)$ denotes the stop-gradient operator, and N' is the number of original samples in a mini-batch. The mini-batch index set is $\mathcal{I} = \{1, \dots, 2N'\}$. Each sample contributes two inputs during training, namely a raw image and a paired segmentation image from the dataset. These inputs yield two disease representations. Hence, the contrastive set contains $2N'$ representations, leading to $I = \{1, \dots, 2N'\}$. The positive set is defined by label-consistent disease representations, not by image pairs themselves. For sample i with disease label y_i , the positive set consists of disease representations from the same disease category. The positive index set is $\mathcal{P}(i) = \{p \in I \mid p \neq i, y_p = y_i\}$ with corresponding disease representations $\{D_p\}_{p \in \mathcal{P}(i)}$. Negatives consist of two parts: (i) disease representations from different disease categories, with index set $\mathcal{N}_d(i) = \{k \in \mathcal{I} \mid y_k \neq y_i\}$ and corresponding disease representations $\{D_k\}_{k \in \mathcal{N}_d(i)}$; and (ii) all crop-domain representations in the mini-batch, where the crop-domain negative index set is $\mathcal{N}_p = \mathcal{I}$ with crop representations $\{P_a\}_{a \in \mathcal{I}}$. By restricting the contrastive anchor to the disease-invariant semantic branch and applying a stop-gradient operation to other vectors, the resulting training signal pulls same-disease samples across crops closer while pushing different diseases farther apart. This mechanism imposes two complementary constraints on the disease semantic branch by separating different diseases while discouraging reliance on domain cues, thereby suppressing semantic–domain information leakage and improving cross-domain generalization.

3.4.3. Domain-Adversarial Semantic-Domain Disentanglement

Semantic–domain disentanglement is promoted from two perspectives, namely semantic anchoring and similarity-based contrastive objectives. However, due to data structure constraints, the PlantVillage training set used here exhibits strong class-conditional domain correlation, because each disease type appears in only two crops. Under this condition, the model may encode crop-domain information into local subspaces or statistics of the disease representation in a bypass manner without substantially disrupting the contrastive similarity structure. Such residual domain signals can remain in dimensions that are relatively insensitive to the contrastive objective. Although these signals may not increase the loss on the seen domains, they can induce a class-conditional distribution shift on unseen crop domains, making the decision boundary sensitive to domain changes and thus weakening generalization. In addition, the training set exhibits strong domain correlation, data imbalance, and limited imaging diversity, which can further exacerbate the class-conditional distribution shift on unseen crop domains and ultimately limit cross-domain generalization. Domain-adversarial learning provides a complementary and stronger constraint by directly minimizing the domain decodability of disease representations. By introducing a crop-domain discriminator on top of disease features and performing min–max optimization via a gradient reversal mechanism, the feature extractor is explicitly driven to produce disease representations from which crop identity is difficult to predict. This optimization-level constraint suppresses any residual domain information that could be exploited by a domain classifier. In effect, domain-adversarial learning explicitly minimizes the separability between disease features and crop-domain labels, imposing an upper

bound on domain information in the disease representation and blocking residual domain signals that remain decodable. Concretely, crop labels are used as domain supervision. A lightweight crop-domain classifier C_c is attached to the disease-branch representation D_i , and a Gradient Reversal Layer is applied to construct a min–max adversarial objective, forcing D_i to be predictive of disease while being uninformative about crop identity. The domain-adversarial loss is defined as

$$\mathcal{L}_{dom}^{grl} = \frac{1}{2N'} \sum_{i \in \mathcal{I}} \text{CE}(C_c(\text{GRL}(D_i)), c_i), \quad (4)$$

where D_i denotes the feature extracted by the disease branch for the i -th sample, c_i is the corresponding crop-domain label, $C_c(\cdot)$ is the lightweight crop-domain discriminator, $\text{CE}(\cdot)$ is the cross-entropy loss. This GRL-based adversarial objective follows the DANN-style training paradigm [36]. It suppresses hidden domain shortcuts from the perspective of decodability without relying on changes in similarity structure, thereby stabilizing class-conditional distributions and decision boundaries on unseen crop domains and improving robustness and generalization for cross-crop disease recognition. The total loss is

$$\mathcal{L}_{total} = \mathcal{L}_{msc} + \lambda_1 (\mathcal{L}_{cos}^d + \mathcal{L}_{cos}^p) + \lambda_2 \mathcal{L}_{dom}^{grl}. \quad (5)$$

3.5. Category-Prototype-Based Fine-Tuning for Classification

During fine tuning, a learnable prototype-based classifier is adopted for disease concept prediction. Owing to the constraints introduced during training, learnable prototypes are updated within a domain-invariant disease semantic space, thereby achieving a better trade-off between discriminability and generalization. Specifically, for each class $c \in \{1, \dots, C\}$, a trainable prototype vector $w_c \in \mathbb{R}^d$ is introduced. The negative squared Euclidean distance between the disease semantic embedding D_i and the prototype is used as the classification logit [37,38]:

$$\ell_{i,c} = -\|D_i - w_c\|_2^2, \quad p(y_i = c \mid x_i) = \frac{\exp(\ell_{i,c})}{\sum_{k=1}^C \exp(\ell_{i,k})}. \quad (6)$$

During training, supervised cross-entropy is minimized to update both the encoder and prototype parameters:

$$\mathcal{L}_{ce} = \frac{1}{N} \sum_{i=1}^N \text{CE}(\text{softmax}(\ell_i), y_i). \quad (7)$$

Learnable prototypes can more readily adapt to calibrate class centers under class imbalance and cross-crop appearance variations, thereby yielding a more stable decision boundary.

3.6. Dataset

Experiments were conducted on the PlantVillage (PV) dataset [39]. PV contains 54,306 leaf images from 38 original crop–disease combination classes. These labels are defined at the crop–disease combination level rather than at the disease-concept level. For cross-crop disease recognition, the original labels were reorganized into a disease-concept label space. Identically named diseases across crops were merged. Healthy was treated as an independent concept. The prediction target is the disease concept, whereas crop labels are used only to define crop domains and evaluation splits.

Only disease concepts appearing in at least two crop domains were retained. This constraint enables cross-crop positive supervision during training. The resulting subset contains 6 disease concepts, including Healthy, across 8 crop domains. It includes

21,521 original images with background and 19,369 segmented images. Segmented images were used only during training. The dataset used in this paper therefore contains both images with background and images without background, as shown in Figure 3. The figure presents representative examples and illustrates the split between seen and unseen domains.

Crop-domain generalization was evaluated under a leave-one-domain-out protocol. One crop domain was treated as unseen, and the remaining domains were used for training and validation. Unseen-domain evaluation was restricted to disease concepts shared by the training domains and the held-out domain. Because the bacterial spot is the only concept present in three crop domains, full rotation over all eight domains is not feasible. Restricted cross-validation was therefore performed by alternating the unseen domain between Peach and Pepper.

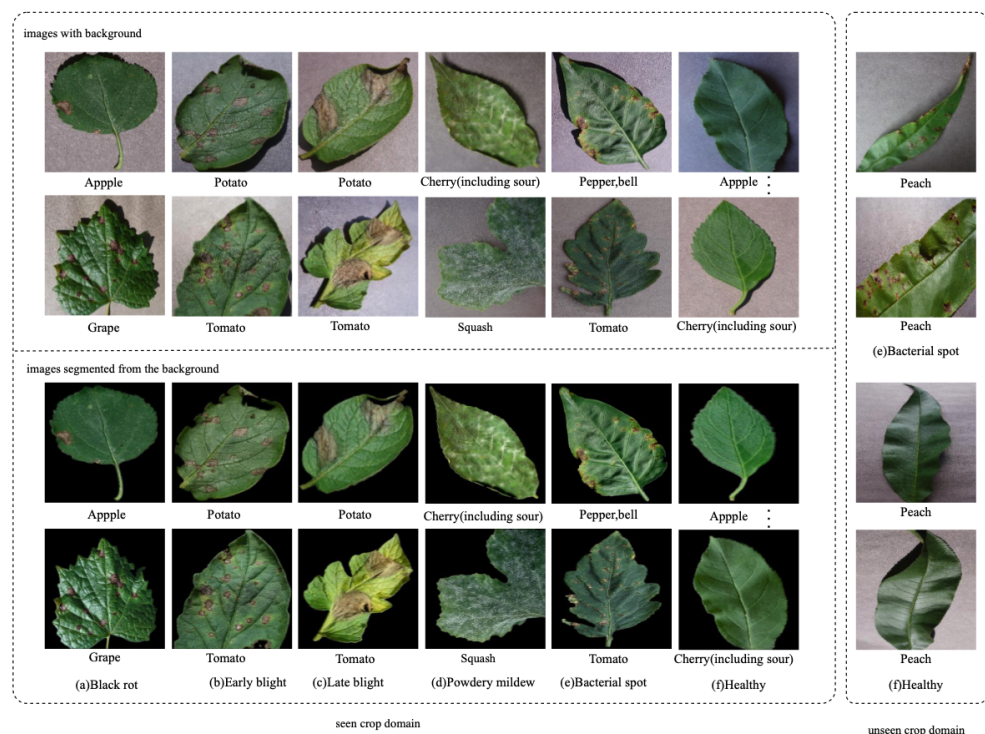


Figure 3. Representative samples from the PlantVillage subset used in this study. Original images with background are shown in the upper part, and the corresponding background-removed images are shown in the lower part. The left side presents samples from the seen crop domains, while the right side presents samples from the unseen crop domain.

Table 1 reports the detailed dataset configuration when Peach is used as the unseen crop domain. Seven crop domains were used for training and validation. One crop domain was held out as a strictly unseen test domain. The held-out domain was not used for model selection. Within the seen domains, samples in each combination of crop domain and disease concept were divided into training and validation sets at a ratio of 9:1. This stratified split preserved the relative class composition within each seen domain. It enabled a more controlled evaluation of in-domain performance and unseen-domain generalization. The retained crop domains are not class-balanced. Apple, Grape, Cherry (including sour), Pepper (bell), and Peach each include 2 classes. Potato includes 3 classes, Tomato includes 4 classes, and Squash includes 1 class. Under the restricted cross-validation protocol, the held-out domains are Peach and Pepper. For each held-out domain, 2 classes are assessed, namely Bacterial spot and Healthy.

To further validate the proposed method, an external strawberry powdery mildew dataset [40] was used. The dataset was obtained from the Strawberry powdery mildew image dataset released on Zenodo. Strawberry was treated as an unseen crop domain and was absent from the PV training data. This dataset was used only for external testing. No model updating was performed on it. To evaluate robustness under low-light conditions, a degraded version of this dataset, denoted as SP_low, was further constructed. The degradation was applied at the pixel level without changing the original geometry or semantic content. It includes reduced brightness, additional noise, a slight color bias, and mild blur. Figure 4 presents representative examples from the real-world strawberry powdery mildew dataset and its low-light counterpart. The figure illustrates the visual characteristics of the external test data used for cross-dataset evaluation.

Table 1. Data distribution in the constrained PlantVillage subset. “With background” denotes original images, and “segmented” denotes background-removed leaf images.

Category	Crop Domain	With Background	Segmented
Bacterial spot	Pepper (bell)	997	897
Bacterial spot	Tomato	2127	1914
Bacterial spot	Peach	2297	0
Black rot	Apple	621	559
Black rot	Grape	1180	1062
Early blight	Potato	1000	900
Early blight	Tomato	1000	900
Late blight	Potato	1000	900
Late blight	Tomato	1909	1718
Powdery mildew	Cherry (including sour)	1052	947
Powdery mildew	Squash	1835	1652
Healthy	Apple	1645	1481
Healthy	Cherry (including sour)	854	769
Healthy	Grape	423	381
Healthy	Pepper (bell)	1478	1330
Healthy	Potato	152	137
Healthy	Tomato	1591	1432
Healthy	Peach	360	0
6	8	21,521	19,369



Figure 4. Strawberry powdery mildew real-world and low-light field datasets.

4. Analysis of Experimental Results

4.1. Experimental Setup and Implementation Details

TDC was implemented under a pretraining–fine-tuning paradigm. The model used a dual-branch visual architecture. The disease branch adopted ViT-B/16 with depth 12. The crop branch adopted ViT-B/16 with depth 6. Both branches were initialized from the pretrained CLIP visual encoder. A frozen CLIP text encoder was used to provide semantic

anchors for disease and crop concepts. This design preserved the general semantic prior learned from large-scale image–text pairs.

During pretraining, all input images were resized to 224×224 pixels. Optimization was performed using AdamW. The initial learning rate was set to 1×10^{-4} . The weight decay was set to 0.05. Training was conducted for 100 epochs. The batch size was set to $E \times 8$, where E denotes the number of crop domains used in training. The temperature parameter in contrastive learning was fixed at 0.07. The text-guidance weight was set to 0.3. The domain-adversarial weight followed a warm-up schedule. It was activated from epoch 9 and increased to 0.35 at epoch 20. To improve reproducibility, deterministic computation was enabled through the corresponding cuDNN settings. The data split seed, denoted as `trial_seed`, was fixed to 0. For each test environment, the data partition remained unchanged, while training was repeated three times with different random seeds (0, 1, and 2). The final result was reported as the average of the three runs.

All experiments were conducted on an NVIDIA RTX 4090 GPU with 24 GB memory. The detailed configuration of the experimental environment is provided in Table 2. At inference time, the text branch was removed, and only the disease branch was retained for classification. This inference protocol is consistent with the task objective, which targets disease recognition rather than crop-domain prediction.

Hyperparameters were optimized using a coarse-to-fine strategy based on validation performance. Initial values and candidate ranges were determined from prior experience, common practice, and preliminary pilot experiments. For the most influential hyperparameters, one-factor analysis was then conducted to further narrow the search space. A targeted grid search was subsequently performed within the reduced ranges, and the final configuration was selected according to the best harmonic mean (HM) on the validation set. Finally, sensitivity analysis was carried out in the neighborhood of the selected optimum to assess the robustness of the configuration. The corresponding search space and selected values are summarized in Table 3.

Table 2. Experimental environment configuration.

Environment	Parameters
GPU	NVIDIA GeForce RTX 4090 (24 G)
CPU	Intel(R) Core(TM) i9-14900K 3.20 GHz
Development	PyCharm 2024.3.4
Language	Python 3.13.2
Framework	PyTorch 2.6.0
Operating platform	CUDA 12.4
Operating System	Windows 11 Professional

Table 3. Hyperparameter search space and selected values.

Hyperparameter	Empirical Initialization	Search Space	Selected Value
Learning rate	1×10^{-4}	$\{5 \times 10^{-5}, 1 \times 10^{-4}, 2 \times 10^{-4}, 5 \times 10^{-4}\}$	1×10^{-4}
Batch size	16	$\{8, 16, 32\}$	8
Temperature τ	0.07	$\{0.03, 0.05, 0.07, 0.10, 0.15\}$	0.07
Text weight	0.3	$\{0.1, 0.3, 0.5, 0.8\}$	0.3
Da weight	0.30	$\{0.25, 0.30, 0.35, 0.40\}$	0.35
Epochs	100	$\{50, 100, 150\}$	100

4.2. Comparative Evaluation

4.2.1. Performance Analysis on Seen and Unseen Domains

To systematically evaluate the effectiveness of the proposed method, comparative experiments are conducted on the PV dataset against eight representative baseline models. The baselines span different learning paradigms to ensure comprehensive and interpretable

comparisons. Specifically, a standard supervised ViT backbone is first included as a reference to characterize the performance ceiling under a purely supervised paradigm as well as the cross-domain generalization bottleneck. Second, the conditional multi-task learning method CMTL-ViT [41] is introduced to assess the extent to which task-conditioning information improves transferability in disease recognition. The third category of methods focuses on a key research direction in the field of unseen crop disease recognition, namely, the representation learning of unseen crop-disease combinations. This category includes the purely visual FF-ViT [22] and its self-supervised variant, CL-ViT [23], with FF-ViT being widely adopted as the baseline in this setting. The fourth category comprises vision–language pretraining and language-augmented visual representation methods, as exemplified by CLIP, CoCoOp [42], and FF-CLIP [32]. In particular, FF-CLIP is specifically designed to classify previously unseen crop–disease combinations. Finally, to validate the applicability of existing domain-generalization paradigms to the specific task of recognizing previously unseen crop diseases, two domain-generalization methods were included for comparison. DADA [24] is a domain-generalization approach based on adversarial disentanglement, whereas DisMAE [26] is a self-supervised reconstruction method that integrates disentangled representation learning. The Top-1 classification accuracy of all models on the seen and unseen domains is reported in Table 4. In addition, the harmonic mean (HM) of these two scores is provided, offering a comprehensive indicator of how well each model balances in-domain performance with out-of-domain generalization ability.

All methods were evaluated under the same benchmark protocol. The compared methods used the same constrained PlantVillage subset, the same disease-concept label space, and the same leave-one-domain-out split. For fair comparison, all methods were implemented with a unified ViT-B/16 backbone. Methods originally designed for unseen crop–disease composition recognition were adapted by changing the output target from crop–disease combinations to disease concepts while preserving their original learning strategies. For supervised and domain-generalization methods, all models were trained on the same source-domain data and evaluated directly on the same held-out crop domain.

The proposed method maintains strong performance on the seen domains while delivering the highest accuracy of 74.29% on the unseen domains and the best HM of 84.53%. Compared with the current strongest baseline, FF-CLIP, its HM is further improved by 2.90 percentage points. These results indicate that the method achieves a better balance between in-domain performance and out-of-domain generalization, demonstrating greater robustness and transferability, particularly for the critical task of recognizing previously unseen crop diseases.

Table 4. Performance comparison on seen and unseen crop domains.

Model	Seen (%)	Unseen (%)	HM (%)
ViT	99.34 ± 0.1	24.78 ± 0.3	39.67 ± 0.3
CMTL-ViT	99.49 ± 0.2	31.72 ± 0.6	48.10 ± 0.6
DADA	97.76 ± 0.3	42.32 ± 0.4	59.07 ± 0.4
FF-ViT	99.53 ± 0.1	52.15 ± 0.3	68.44 ± 0.3
CL-ViT	99.31 ± 0.3	54.28 ± 0.4	70.19 ± 0.4
CLIP	98.59 ± 0.2	49.32 ± 0.3	65.75 ± 0.3
CoCoOp	99.36 ± 0.5	55.06 ± 0.6	70.86 ± 0.6
FF-CLIP	99.19 ± 0.2	69.35 ± 0.5	81.63 ± 0.4
DisMAE	98.56 ± 0.4	68.27 ± 0.8	80.67 ± 0.8
TDC	98.04 ± 0.4	74.29 ± 0.7	84.53 ± 0.6

4.2.2. Parameter Efficiency and Inference Overhead

To evaluate deployment friendliness from the perspective of parameter efficiency, the total number of inference-time parameters and the average per-image inference latency were further compared. The results are summarized in Table 5. The comparison was conducted on a representative subset of disease-recognition methods under the same measurement setting. Within this subset, the proposed method uses the fewest inference-time parameters, achieves the lowest inference latency, and attains the highest HM. These results suggest a favorable efficiency–accuracy trade-off among the evaluated methods. In addition to its strong performance on unseen domains, this method maintains a relatively balanced profile in terms of model efficiency and predictive accuracy.

Table 5. Comparison of parameter efficiency and inference overhead across models.

Model	CMTL-ViT	FF-ViT	CL-ViT	FF-CLIP	TDC
Total parameters (M)	89	200	125	310	85
Execution time (ms)	1.91	4.69	3.43	4.97	1.89
HM (%)	48.10	68.44	70.19	81.63	84.53

4.2.3. Cross-Dataset Generalization and Distribution Robustness Evaluation

Due to data structure limitations, the unseen domain evaluation on the PV dataset, excluding the healthy category, is only validated on the unseen crop–disease combination of Peach and Bacterial spot. The unseen domain accuracy for this combination is 71.27%. To further eliminate model overfitting to specific crops or data distributions, and to test its cross-domain generalization ability, we conducted cross-dataset testing on datasets with significant differences in acquisition conditions and data distributions. We evaluated the model on the PlantDoc [43] dataset and a strawberry powdery mildew dataset collected from actual field conditions, where strawberry was an unseen crop during training. Additionally, we constructed a low-light test set, SP_low, based on the SP dataset. SP_low consists of normally exposed images, simulating low-light imaging characteristics without altering the geometric structure or semantic content. This includes reduced overall brightness, a lower signal-to-noise ratio, and slight color bias with mild blur, which were used to analyze the model’s robustness and scalability under low-light conditions. The assessments reported in this subsection do not belong to a single evaluation category. Rather, they represent several distinct but complementary forms of evidence. The PV result corresponds to in-benchmark unseen-domain evaluation. The PlantDoc result reflects robustness to cross-source distribution shift. The SP result provides evidence of cross-crop transferability on a real-field dataset. The SP_low result assesses robustness under low-light image degradation. These results should therefore be interpreted as complementary rather than homogeneous. Taken together, they indicate that the learned representation remains relatively stable under multiple forms of distribution shift. The results are shown in Table 6.

Table 6. Evaluation on real-world and low-light datasets.

Dataset	Category	Unseen (%)
PV	Bacterial spot	71.27
PlantDoc	Bacterial spot	61.11
SP	Powdery mildew	81.99
SP_low	Powdery mildew	76.31

In the PlantDoc dataset, the crop categories overlap highly with the seen crop domain in the PV dataset, making it challenging to create a completely consistent unseen crop domain as in PV. However, there are significant differences in the imaging devices, shooting

distance, and background between PlantDoc and PV, which allows for the evaluation of the model's robustness to distribution shifts across data sources. The results show that the model achieves an accuracy of 61.11% on PlantDoc, suggesting that the learned representation retains a certain degree of cross-source transferability.

To examine whether the learned representation extends beyond a single unseen crop–disease case, additional evaluation was conducted on the SP and SP_low datasets. The SP dataset contains field-acquired images of strawberry powdery mildew, where strawberry was not observed during training. Under this setting, the method achieves an accuracy of 81.99%. The SP_low dataset was constructed from SP to simulate low-light degradation while preserving the geometric structure and semantic content of the original images. On this set, the accuracy is 76.31%. The decrease from SP to SP_low is limited in magnitude, which is consistent with a degree of tolerance to the constructed illumination-related perturbation. Accordingly, the SP and SP_low results are more appropriately interpreted as suggestive evidence of transferability and low-light tolerance.

In summary, although the unseen-domain evaluation on PV is limited to a single unseen crop–disease combination, the additional evaluations under different conditions provide complementary evidence relevant to generalization beyond the benchmark setting. Taken together, these results indicate that the learned representation retains a degree of transferability under the tested unseen or shifted conditions and shows some tolerance to low-light degradation.

4.3. Ablation Studies

To better reflect the training pipeline, the ablation analysis considers both the pre-training stage and the fine-tuning stage. Table 7 reports the effect of the main pretraining components in terms of HM. Starting from the dual-branch backbone, adding contrastive learning alone yields only a limited gain, increasing HM from 74.71% to 75.19%. In contrast, text guidance and domain-adversarial learning each lead to more noticeable improvements, reaching 79.64% and 77.89%, respectively. When these components are combined, the gain becomes more pronounced, and the full pretraining configuration achieves the best result of 83.39%. These results indicate that semantic guidance and domain-related suppression contribute in a complementary manner during pretraining.

Table 7. Pretraining ablation.

Dual-Branch Backbone	Contrastive	Text Guidance	Domain Adversarial	HM (%)
✓				74.71 ± 0.3
✓	✓			75.19 ± 0.5
✓		✓		79.64 ± 0.6
✓			✓	77.89 ± 0.9
✓	✓	✓		81.84 ± 0.6
✓	✓		✓	82.12 ± 0.8
✓		✓	✓	81.62 ± 0.5
✓	✓	✓	✓	83.39 ± 0.6

✓ indicates that the corresponding component is enabled.

Table 8 further examines the effect of prototype-based fine-tuning. On top of the full pretraining configuration, adding the prototype constraint improves HM from 83.39% to 84.53%. This result suggests that prototype regularization provides an additional benefit beyond pretraining alone. Overall, the gains of TDC are not driven by any single component; rather, they arise from the coordinated optimization across three aspects, namely semantic alignment, domain invariance, and intra-class aggregation, thereby enabling stronger generalization in unseen-crop disease recognition.

Table 8. Prototype ablation.

Setting	HM (%)
Without prototype	83.39 ± 0.7
With prototype	84.53 ± 0.6

To justify the prompt design in Section 3.4.1, an additional comparison of text templates was conducted. Three prompt forms were considered: (1) the category-name template used in the main model, (2) a manually written descriptive template, and (3) a decoupled descriptive template. The category-name template uses only the disease or crop label, such as “a photo of a disease”. The descriptive template expands the prompt with textual attribute descriptions. The decoupled descriptive template further separates disease-related and crop-related textual content. All variants were evaluated under the same training protocol and the same unseen-domain test setting. The objective was to examine whether more detailed textual descriptions provide stronger semantic guidance than the simple label-based template.

As shown in Table 9, the category-name template yields the best performance on both the seen and unseen domains. The gain on the seen domain is marginal. The gain on the unseen domain is more pronounced. Compared with the descriptive template, unseen accuracy increases from 69.85% to 74.29%. Compared with the decoupled descriptive template, unseen accuracy increases from 72.21% to 74.29%. The results indicate that under domain shift, the decoupled descriptive template further reduces semantic entanglement relative to the descriptive design. However, its performance remains inferior to that of the concise category-name template. Concise label-level prompts provide more stable semantic anchors for cross-modal alignment.

To assess whether the reported performance depends on an overly narrow parameter configuration, sensitivity analysis was conducted for the most influential hyperparameters. The analysis focuses on the text-alignment loss weight and the domain-adversarial loss weight, as these terms directly control the strength of semantic anchoring and domain-invariant representation learning, which are central to the proposed framework. The final training configuration is given in Section 4.1. In each analysis, only one hyperparameter was varied, while all others were fixed to the final selected configuration. Performance was evaluated on both the seen and unseen domains. The harmonic mean (HM) of the two Top-1 accuracies was used as the main summary metric. As shown in Table 10, the model remains relatively stable under moderate perturbations around the selected values.

Table 9. Effect of different prompt templates on seen-domain and unseen-domain classification performance.

Prompt Type	Seen (%)	Unseen (%)
Descriptive template	97.97	69.85
Decoupled descriptive template	98.00	72.21
Category-name template (ours)	98.04	74.29

Table 10. Sensitivity analysis of key hyperparameters measured by HM (%).

Hyperparameter Setting	Value	HM (%)
Text weight	0.1	82.27
Text weight	0.3	84.53
Text weight	0.5	83.58
Da weight	0.30	83.61
Da weight	0.35	84.53
Da weight	0.40	80.63

4.4. Representation Visualization

To provide an intuitive assessment of whether the disease representations learned by TDC are consistent across crop domains, t-SNE was applied to visualize the features output by the disease branch, as shown in Figure 5. The visualization suggests that the learned disease representations are relatively more aligned with pathology-related cues than with crop-domain-specific appearance factors. A similar cross-domain mixing pattern is observed for the Healthy class with a comparatively stable intra-class structure. These patterns are consistent with the interpretation that the proposed method may capture cross-domain disease semantics and reduce reliance on superficial crop-domain cues. This visualization provides qualitative support for this interpretation.

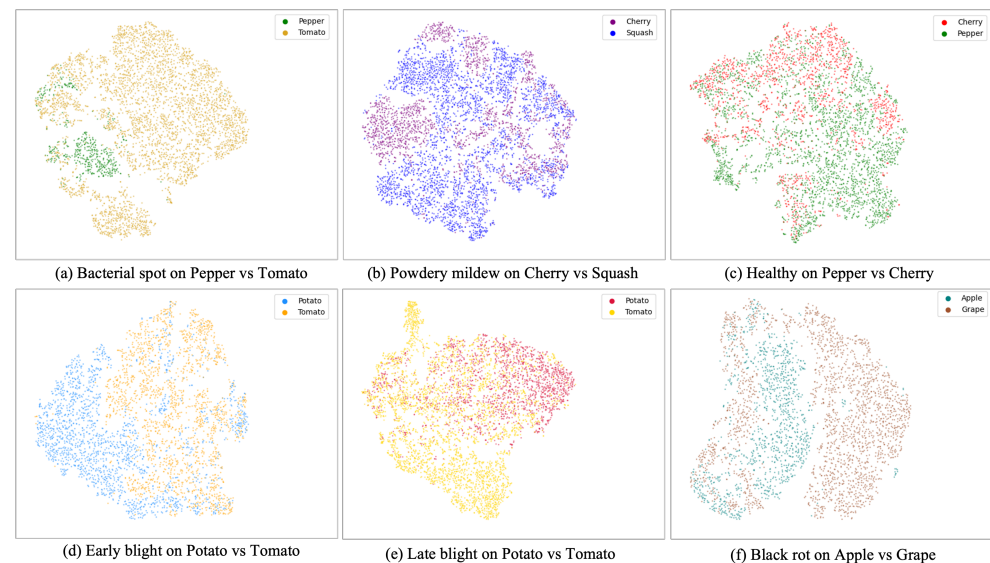


Figure 5. t-SNE visualization of the domain-invariant disease representations learned by TDC on the dataset. Each subplot shows the embedding distribution of a shared cross-domain category under different crop domains with point colors indicating the crop-domain source.

4.5. Semantic Invariance and Disentanglement Verification

In unseen-crop disease classification, the domain invariance of the learned disease representations is assessed using the crop-domain decodability of the disease-branch features. Specifically, a crop-species classifier is trained on the disease-branch features, and its accuracy reflects the amount of residual crop-domain information encoded in the disease representation. If this branch captures disease semantics that are independent of the crop domain, its ability to discriminate crop species should be close to random; under this setting, the random baseline is 14.29%. Incorporating text guidance and contrastive constraints improves feature discriminability and yields a certain degree of disentanglement. However, because crop appearance and disease symptoms are strongly correlated visually, the crop-domain classification accuracy based on disease features remains persistently high, indicating that domain-related factors have not been sufficiently removed. This phenomenon can be attributed to the strong visual coupling between crop appearance and symptom patterns. In contrast, after feature learning becomes relatively stable, introducing GRL-based domain-adversarial training further suppresses and removes residual crop-domain cues while preserving domain-invariant disease semantics. Correspondingly, the crop-domain classification accuracy drops substantially and gradually stabilizes, indicating that domain-adversarial learning further encourages the model to capture disease semantic features with stronger cross-domain consistency. Taken together, these results provide supportive evidence that the learned disease representations become less crop-domain-

decodable while retaining disease-relevant semantics. This pattern is consistent with improved representational generalization in cross-crop recognition, as shown in Figure 6.

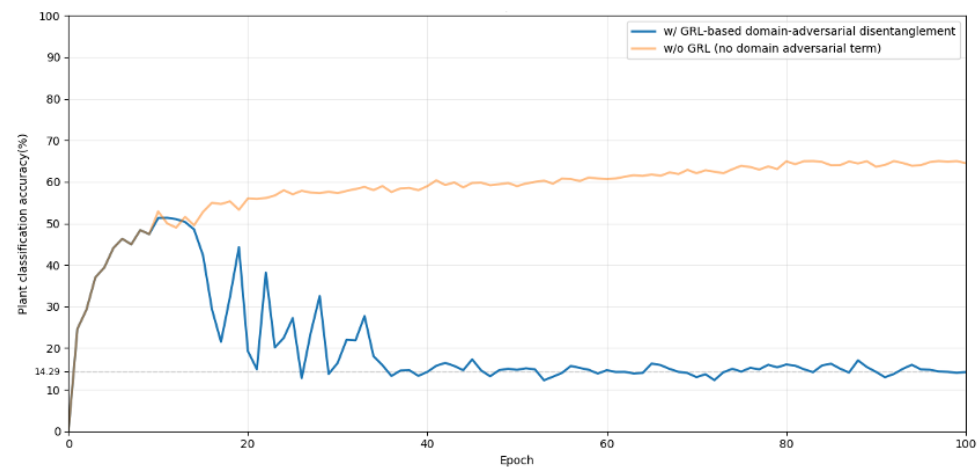


Figure 6. Domain-disentanglement evaluation for disease representations. A crop-domain classifier is trained on features produced by the disease branch, and its classification accuracy serves as a proxy for crop-domain information leakage. The orange curve reports crop-domain accuracy when only text guidance and mixed label-space contrastive learning are applied. The blue curve reports crop-domain accuracy after incorporating domain-adversarial training using a gradient reversal layer. The gray dotted line represents the random baseline for classification.

5. Discussion

Compared with existing supervised, disentanglement-based, and vision–language methods, the present framework is better suited to the task of recognizing seen diseases in unseen crop domains. The results suggest that the performance gain does not arise only from stronger visual encoding. A more important factor lies in how disease semantics and crop-domain variation are organized during representation learning. Compared with conventional supervised methods and domain-generalization baselines, the proposed design is more consistent with the current task setting, which requires the recognition of seen diseases in unseen crop domains. Purely visual disentanglement or reconstruction-based methods can improve robustness. However, they do not explicitly anchor disease features to a cross-domain semantic reference. Vision–language baselines introduce semantic priors. However, they still do not sufficiently constrain the leakage of crop-domain information into disease representations. From the deployment perspective, retaining only the disease branch during inference enables strong parameter efficiency and low latency. This makes the method suitable for field inspection and early warning applications. It also supports its use as a core visual module in multimodal agricultural perception systems. The method does not require attribute annotation. Instead, it uses a text-guided disentanglement mechanism to separate disease semantics from crop-domain-related factors with the aim of learning disease-invariant features that remain robust across plant species. To avoid ambiguity, the claimed state-of-the-art performance is limited to the setting of unseen crop-domain generalization.

Several limitations still remain in this paper. Owing to constraints in the dataset structure, only disease concepts appearing in at least two crop domains were retained. As a result, the final benchmark covers only six disease concepts and eight crop domains. Therefore, the current evaluation does not yet cover a broader range of crop species and disease categories. This choice helps the model focus on common patterns across plants and allows a clearer validation of methodological feasibility. However, it also limits applicability to more diverse plant species and disease types. Therefore, the current benchmark

is better interpreted as a controlled feasibility test rather than a fully comprehensive cross-crop benchmark.

Accordingly, future work will further expand the study to larger real-world field datasets that cover multiple dimensions, including species, season, region, and device conditions. More plant species and disease categories will also be included to improve coverage and practical applicability. The current setting assumes that all disease concepts are already seen. In real deployment, however, the system may face novel diseases, co-infections, and differences in disease severity. The integration of open-set recognition and uncertainty estimation may improve operational safety. The text-guided mechanism may also be further optimized to improve accuracy and usability under diverse plant backgrounds. Another important direction is the construction of cross-crop datasets with attribute annotations and compositional graphs together with the exploration of integrating TDC as a disease-semantic disentanglement module into compositional zero-shot learning frameworks. Embedding the framework into IoT and edge-computing environments for near-real-time monitoring, and exploring multimodal fusion with spectral images, weather signals, and agronomic sensor data, may further improve robustness under extreme environmental changes such as unseen crops. In addition, uncertainty estimation and active learning may help reduce annotation cost and adaptation overhead during deployment.

6. Conclusions

Motivated by the practical need for transferring the recognition of the same disease across crop species in real agricultural settings, this paper proposes TDC from a domain-generalization perspective. The framework learns cross-crop transferable disease representations through three complementary elements: semantic anchoring, structural disentanglement, and optimization constraints, thereby improving stability in cross-crop disease classification. Specifically, a semantic anchor is introduced to guide visual disease features toward a semantically invariant space, and an anchor-driven contrastive disentanglement strategy is adopted to effectively separate disease-related cues from plant-related factors. In addition, domain-adversarial learning via a Gradient Reversal Layer (GRL) and prototypical metric-based classification are incorporated to enhance the domain indiscriminability of disease representations and to improve the discriminative stability for key categories. Experimental results show that the proposed method outperforms existing state-of-the-art models on unseen crop-domain disease classification, substantially improving cross-crop generalization while preserving strong performance on seen-domain tasks. Beyond performance gains, the framework provides a reproducible, interpretable, and deployable solution for cross-crop disease recognition, and it demonstrates the methodological value of using vision–language concept priors as semantic anchors to drive disentanglement-based domain generalization in agricultural intelligent perception. These findings establish a solid foundation for the rapid deployment of disease early-warning systems to new crop species.

Author Contributions: Conceptualization, Z.W., J.G. and H.J.; methodology, J.G.; software, W.H.; validation, K.C.; formal analysis, Z.W.; investigation, K.C.; resources, Z.W.; data curation, K.Z.; writing—original draft preparation, Z.W., J.G., W.H., K.Z. and K.C.; writing—review and editing, Z.W., J.G. and H.J.; visualization, K.Z.; supervision, H.J.; project administration, H.J.; funding acquisition, Z.W. and H.J. All authors have read and agreed to the published version of the manuscript.

Funding: This paper was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP)—Innovative Human Resource Development for Local Intellectualization program grant funded by the Korean government (MSIT) (IITP-2026-RS-2022-00156334, contribution rate: 70%) and by the Doctoral Fund Project of Weifang University of Science and Technology (KJRC2023045, contribution rate: 30%).

Institutional Review Board Statement: Not applicable.

Data Availability Statement: This paper analyzed the publicly available PlantVillage dataset and selected a subset of the data for experiments.

Acknowledgments: The authors would like to express their gratitude to all the contributors to this paper.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Savary, S.; Willocquet, L.; Pethybridge, S.J.; Esker, P.; McRoberts, N.; Nelson, A. The global burden of pathogens and pests on major food crops. *Nat. Ecol. Evol.* **2019**, *3*, 430–439. [\[CrossRef\]](#)
2. Gai, Y.; Wang, H. Plant disease: A growing threat to global food security. *Agronomy* **2024**, *14*, 1615. [\[CrossRef\]](#)
3. Bock, C.H.; Barbedo, J.G.; Del Ponte, E.M.; Bohnenkamp, D.; Mahlein, A.K. From visual estimates to fully automated sensor-based measurements of plant disease severity: Status and challenges for improving accuracy. *Phytopathol. Res.* **2020**, *2*, 9. [\[CrossRef\]](#)
4. Upadhyay, A.; Chandel, N.S.; Singh, K.P.; Chakraborty, S.K.; Nandede, B.M.; Kumar, M.; Subeesh, A.; Upendar, K.; Salem, A.; Elbeltagi, A. Deep learning and computer vision in plant disease detection: A comprehensive review of techniques, models, and trends in precision agriculture. *Artif. Intell. Rev.* **2025**, *58*, 92. [\[CrossRef\]](#)
5. Ebrahimi, M.A.; Khoshtaghaza, M.H.; Minaei, S.; Jamshidi, B. Vision-based pest detection based on SVM classification method. *Comput. Electron. Agric.* **2017**, *137*, 52–58. [\[CrossRef\]](#)
6. Ashurov, A.Y.; Al-Gaashani, M.S.A.M.; Samee, N.A.; Alkanhel, R.; Atteia, G.; Abdallah, H.A.; Saleh Ali Muthanna, M. Enhancing plant disease detection through deep learning: A Depthwise CNN with squeeze and excitation integration and residual skip connections. *Front. Plant Sci.* **2025**, *15*, 1505857. [\[CrossRef\]](#)
7. George, R.; Thuseethan, S.; Ragel, R.G.; Mahendrakumaran, K.; Nimishan, S.; Wimalasooriya, C.; Alazab, M. Past, present and future of deep plant leaf disease recognition: A survey. *Comput. Electron. Agric.* **2025**, *234*, 110128. [\[CrossRef\]](#)
8. Mohanty, S.P.; Hughes, D.P.; Salathé, M. Using deep learning for image-based plant disease detection. *Front. Plant Sci.* **2016**, *7*, 215232. [\[CrossRef\]](#)
9. Karthikeyan, M.; Raja, D. Deep transfer learning enabled DenseNet model for content based image retrieval in agricultural plant disease images. *Multimed. Tools Appl.* **2023**, *82*, 36067–36090. [\[CrossRef\]](#)
10. Eunice, J.; Popescu, D.E.; Chowdary, M.K.; Hemanth, J. Deep learning-based leaf disease detection in crops using images for agricultural applications. *Agronomy* **2022**, *12*, 2395. [\[CrossRef\]](#)
11. Zhou, K.; Liu, Z.; Qiao, Y.; Xiang, T.; Loy, C.C. Domain generalization: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 4396–4415. [\[CrossRef\]](#)
12. Wang, X.; Chen, H.; Tang, S.A.; Wu, Z.; Zhu, W. Disentangled representation learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 9677–9696. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Shoaib, M.; Shah, B.; Ei-Sappagh, S.; Ali, A.; Ullah, A.; Alenezi, F.; Gechev, T.; Hussain, T.; Ali, F. An advanced deep learning models-based plant disease detection: A review of recent research. *Front. Plant Sci.* **2023**, *14*, 1158933. [\[PubMed\]](#)
14. Wu, X.; Fan, X.; Luo, P.; Choudhury, S.D.; Tjahjadi, T.; Hu, C. From laboratory to field: Unsupervised domain adaptation for plant disease recognition in the wild. *Plant Phenomics* **2023**, *5*, 0038. [\[CrossRef\]](#) [\[PubMed\]](#)
15. Gao, X.; Feng, Q.; Wang, S.; Zhang, J.; Yang, S. A multi-source domain feature adaptation network for potato disease recognition in field environment. *Front. Plant Sci.* **2024**, *15*, 1471085. [\[CrossRef\]](#)
16. Yang, S.; Feng, Q.; Zhang, J.; Yang, W.; Zhou, W.; Yan, W. From laboratory to field: Cross-domain few-shot learning for crop disease identification in the field. *Front. Plant Sci.* **2024**, *15*, 1434222. [\[CrossRef\]](#)
17. Zhan, K.; Peng, Y.; Liao, M.; Wang, Y. Domain generalization plant leaf disease recognition: Toward from laboratory to field. *Eng. Appl. Artif. Intell.* **2025**, *156*, 111168. [\[CrossRef\]](#)
18. Bouacida, I.; Farou, B.; Djakhdjakha, L.; Seridi, H.; Kurulay, M. Innovative deep learning approach for cross-crop plant disease detection: A generalized method for identifying unhealthy leaves. *Inf. Process. Agric.* **2025**, *12*, 54–67. [\[CrossRef\]](#)
19. Kumar, P.; Mathew, J.; Sanodiya, R.K.; Setty, T.; Bhaskarla, B.P. Zero shot plant disease classification with semantic attributes. *Artif. Intell. Rev.* **2024**, *57*, 305 [\[CrossRef\]](#)
20. Zhang, T.; Liang, K.; Du, R.; Sun, X.; Ma, Z.; Guo, J. Learning invariant visual representations for compositional zero-shot learning. In *European Conference on Computer Vision*; Springer Nature: Cham, Switzerland, 2022; pp. 339–355
21. Lu, X.; Guo, S.; Liu, Z.; Guo, J. Decomposed soft prompt guided fusion enhancing for compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2023; pp. 23560–23569.
22. Chai, A.Y.H.; Lee, S.H.; Tay, F.S.; Then, Y.L.; Goëau, H.; Bonnet, P.; Joly, A. Pairwise feature learning for unseen plant disease recognition. In *2023 IEEE International Conference on Image Processing (ICIP)*; IEEE: Piscataway, NJ, USA, 2023; pp. 306–310.

23. Chai, A.Y.H.; Lee, S.H.; Tay, F.S.; Bonnet, P.; Joly, A. Beyond supervision: Harnessing self-supervised learning in unseen plant disease recognition. *Neurocomputing* **2024**, *610*, 128608. [\[CrossRef\]](#)
24. Peng, X.; Huang, Z.; Sun, X.; Saenko, K. Domain agnostic learning with disentangled representations. In Proceedings of the 36th International Conference on Machine Learning, PMLR 97, Long Beach, CA, USA, 9–15 June 2019; pp. 5102–5112.
25. Lin, W.; Chu, J.; Leng, L.; Miao, J.; Wang, L. Feature disentanglement in one-stage object detection. *Pattern Recognit.* **2024**, *145*, 109878. [\[CrossRef\]](#)
26. Zhang, A.; Wang, H.; Wang, X.; Chua, T.S. Disentangling Masked Autoencoders for Unsupervised Domain Generalization. In *European Conference on Computer Vision*; Springer Nature: Cham, Switzerland, 2024; pp. 126–151.
27. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 8748–8763.
28. Huang, H.; Xia, Y.; Zhou, S.; Wang, H.; Wang, S.; Zhao, Z. Bridging domain generalization to multimodal domain generalization via unified representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: Piscataway, NJ, USA, 2025; pp. 22488–22498.
29. Jiang, J.; He, Z.; Wan, A.; Al-Bukhaiti, K.; Wang, K.; Zhu, P.; Cheng, X. Zero-Shot Industrial Anomaly Detection via CLIP-DINOv2 Multimodal Fusion and Stabilized Attention Pooling. *Electronics* **2025**, *14*, 4785. [\[CrossRef\]](#)
30. Zi, X.; Wu, C. DE-CLIP: Unsupervised Dense Counting Method Based on Multimodal Deep Sharing Prompts and Cross-Modal Alignment Ranking. *Electronics* **2025**, *14*, 1234. [\[CrossRef\]](#)
31. Phan, V.M.H.; Xie, Y.; Qi, Y.; Liu, L.; Liu, L.; Zhang, B.; Liao, Z.; Wu, Q.; To, M.S.; Verjans, J.W. Decomposing disease descriptions for enhanced pathology detection: A multi-aspect vision-language pre-training framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2024; pp. 11492–11501.
32. Liaw, J.Z.; Chai, A.Y.H.; Lee, S.H.; Bonnet, P.; Joly, A. Can Language Improve Visual Features For Distinguishing Unseen Plant Diseases? In *International Conference on Pattern Recognition*; Springer Nature: Cham, Switzerland, 2024; pp. 296–311.
33. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16×16 words: Transformers for image recognition at scale. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 3–7 May 2021.
34. Ma, W.; Li, S.; Zhang, J.; Liu, C.H.; Kang, J.; Wang, Y.; Huang, G. Borrowing knowledge from pre-trained language model: A new data-efficient visual learning paradigm. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: Piscataway, NJ, USA, 2023; pp. 18786–18797.
35. Chen, H.; Zhang, Q.; Huang, Z.; Wang, H.; Zhao, J. Towards domain-specific features disentanglement for domain generalization. *arXiv* **2023**, arXiv:2310.03007. [\[CrossRef\]](#)
36. Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; March, M.; Lempitsky, V. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.* **2016**, *17*, 1–35.
37. Movshovitz-Attias, Y.; Toshev, A.; Leung, T.K.; Ioffe, S.; Singh, S. No fuss distance metric learning using proxies. In *Proceedings of the IEEE International Conference on Computer Vision*; IEEE: Piscataway, NJ, USA, 2017; pp. 360–368.
38. Snell, J.; Swersky, K.; Zemel, R. Prototypical networks for few-shot learning. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30.
39. Hughes, D.; Salathé, M. An open access repository of images on plant health to enable the development of mobile disease diagnostics. *arXiv* **2015**, arXiv:1511.08060.
40. Wang, J. Strawberry Powdery Mildew Image Dataset. *Zenodo* **2025**. [\[CrossRef\]](#)
41. Lee, S.H.; Goëau, H.; Bonnet, P.; Joly, A. Conditional multi-task learning for plant disease identification. In *2020 25th International Conference on Pattern Recognition (ICPR)*; IEEE: Piscataway, NJ, USA, 2021; pp. 3320–3327.
42. Zhou, K.; Yang, J.; Loy, C.C.; Liu, Z. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2022; pp. 16816–16825.
43. Singh, D.; Jain, N.; Jain, P.; Kayal, P.; Kumawat, S.; Batra, N. PlantDoc: A dataset for visual plant disease detection. In *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*; Association for Computing Machinery: New York, NY, USA, 2020; pp. 249–253.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.