IMPROVED CROP DISEASE DETECTION USING HYBRID VISION TRANSFORMERS AND KNOWLEDGE DISTILLATION

FENG JING

UNIVERSITI SAINS MALAYSIA

2025

IMPROVED CROP DISEASE DETECTION USING HYBRID VISION TRANSFORMERS AND KNOWLEDGE DISTILLATION

by

FENG JING

Thesis submitted in fulfilment of the requirements
for the degree of
Doctor of Philosphy

September 2025

ACKNOWLEDGEMENT

I would like to express my deepest gratitude to my main supervisor, **Dr. Ong Wen Eng**, for her invaluable guidance, steadfast support, and unwavering encouragement throughout my research journey. Her timely and insightful feedback, combined with her rigorous attention to detail, provided clarity at every stage and elevated the quality of this work. Her patient mentorship, marked by an exceptional commitment to academic excellence, has not only shaped the direction of this research but has also instilled in me the value of diligence and precision, ensuring that this work meets the highest scholarly standards. I am equally grateful to my co-supervisor, **Associate Professor Dr. Teh Wen Chean**, for his guidance in defining the overarching direction of my research and in making critical decisions that shaped the focus of my work. His insights provided clarity and perspective, allowing me to approach my research with a more strategic and purposeful outlook.

I extend my heartfelt appreciation to the faculty and staff at the School of Mathematics for fostering a supportive and enriching academic environment that has been instrumental to both my research and personal growth. I am especially grateful for the comfortable study spaces provided for graduate students, creating an ideal setting for focused research. Additionally, I am thankful to my fellow students, whose camaraderie and willingness to help have made this journey both productive and enjoyable.

I am deeply grateful to my family, friends, and peers for their unwavering support, patience, and encouragement. Their insightful discussions and moral support have been my strength, enriching my research experience and making this journey meaningful and fulfilling.

My heartfelt thanks go to everyone who contributed to this work, each playing a vital role in shaping its direction and quality. Thank you all for being an essential part of this journey.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF SYMBOLS

α_{ij}	Attention weights
β	Learnable parameters
b	Bias of linear projection
C	Channels of feature maps
$C^{'}$	Channels of the student model's feature maps
C_{in}	Input channel
C_{out}	Output channel
c	Class token
D	Embedding dimension of the model
d_k	The dimension of the input vector
E	Weight matrix
$f_{\mathbf{ap}}$	Average pooling
f^s	Feature vector of the student model
f^t	Feature vector of the teacher model
FC	Fully linear operator
F^S	Feature map of the student model
F^T	Feature map of the teacher model
H	Height of the input image
H'	Height of the feature image
h	Height of the patch window
K	Match query to calculate attention
$\mathcal{L}_{KL}$	Kullback-Leibler divergence

$\mathcal{L}$	Loss function
$\mathcal{L}_{CE}$	Cross-entropy loss function
$\mathcal{L}_{Seg}$	Segmentation task loss
$\mathcal{L}_{TaT}$	Loss of target-aware transformer algorithm to distill knowledge
$\mathcal{L}_{TaT}^{PM}$	Patch-group distillation loss
$\mathcal{L}_{TaT}^{SW}$	Shifted window patch-group distillation loss
$\mathcal{L}_{Task}$	Any loss function applicable to the generic machine learning tasks
M	Window size of patch-group distillation
N	Numbers of patches
n	The teacher model's feature map height
P	Size of the patch
Q	A matrix want to calculate attention
s	Stride of the DWTE
$s^{'}$	Normalization of weights
S	Kernel size of the DWTE
$\tau(\cdot)$	Group-wise
$\Gamma(\cdot)$	Transform 3D feature map into 2D feature map
u	Coarse-grained region pruning coefficient
μ	Mean
V	The actual data of the elements in the sequence
ω	Weight vector of the coarse regions
W	Width of the input image
W'	Width of the feature image
W^k	Weight matrix of key

W^o	Weight matrix applied to the concatenated output of the multi-head attention mechanism
W^q	Weight matrix of query
W^v	Weight matrix of value
W^i	Weights matrix of the student feature
w	Width of the patch window
x_i	The i th patch
$\tilde{X}$	Coarse region
X^r_{avg}	Average pooling point of the coarse region
X^r	Splitted region map
$\hat{z}_i$	Output of layer normalization
Z	The embedded vector sequence
Z_{att}	The output of Multi-Head Attention
z_i	The i th embedded vector
z_i^{pos}	The i th embedded vector with position information
$\theta(\cdot)$	Linear transformation
ζ	Weight factor of the segmentation task loss
σ^2	Variance
$\delta(\cdot)$	Step function
$\langle \cdot, \cdot \rangle$	Inner product

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
CFANet	Coarse-to-Fine Attention Network
CNNs	Convolutional Neural Networks
CV	Computer Vision
DCTN	Dense CNNs and Transformer Network
DeiT	Data-efficient Image Transformers
DSC	Depthwise Separable Convolutions
DWTE	Depth-Wise Split Token Embedding
EGWT	EfficientNet Convolutional Group-Wise Transformer
FLOPs	Floating Point Operations per Second
FN	False Negatives
FP	False Positives
ICVT	Inception and Vision Transformer
IoU	Intersection over Union
ISVRC	ImageNet Large Scale Visual Recognition Challenge
KD	Knowledge Distillation
LN	Layer Normalization
MA-CNN	Multi-Attention Convolutional Neural Network
mIoU	mean Intersection over Union
MLP	Multi-Layer Perceptron
MHA	Multi-Head Attention
NLP	Natural Language Processing

RA-CNN	Recurrent Attention Convolutional Neural Network
SA	Self-Attention
SMOTE	Synthetic Minority Over-sampling Technique
SRCNN	Super Resolution Convolutional Neural Network
SOTA	State Of The Art
Swin	Shifted Window
TaT	Target-aware Transformer
TN	True Negatives
TP	True Positive
ViT	Vision Transformer

PENAMBAHBAIKAN PENGESANAN PENYAKIT TANAMAN MENGGUNAKAN TRANSFOMER PENGLIHATAN HIBRID DAN PENYULINGAN PENGETAHUAN

ABSTRAK

Dalam bidang penglihatan komputer, seni bina Transformer telah muncul sebagai alternatif yang berkuasa kepada rangkaian neural konvolusional tradisional, dengan menawarkan keupayaan pemodelan global yang unggul serta prestasi cemerlang dalam pelbagai tugas pengecaman visual. Walau bagaimanapun, saiz parameter yang besar serta kebergantungan terhadap data latihan berskala besar, di samping kekurangan struktur cekap bagi pengekstrakan ciri berbutir halus, telah mengehadkan penggunaannya dalam aplikasi seperti pengesanan penyakit tanaman yang tepat. Dalam bidang penyulingan pengetahuan, kerangka guru–pelajar konvensional sering menghadapi isu ketidakpadanan antara peta ciri pertengahan akibat perbezaan kedalaman rangkaian. Tambahan pula, dimensi peta ciri pertengahan yang tinggi menyebabkan peningkatan kos pengiraan secara kuadratik apabila penyulingan dilakukan secara langsung, sekali gus mengurangkan kecekapan keseluruhan. Untuk mengatasi cabaran ini, kajian ini terlebih dahulu memperbaiki mekanisme perhatian kendiri berbilang kepala dalam Transformer dengan memperkenalkan reka bentuk perhatian mengikut kumpulan. Pengiraan perhatian dibahagikan kepada beberapa subkumpulan, di mana parameter dikongsi dan perhatian dikira secara bebas. Hasil setiap kumpulan kemudian digabungkan untuk mengekalkan keupayaan pemodelan global sambil mengurangkan bilangan parameter dan beban pengiraan. Berdasarkan mekanisme ini, satu model ringan untuk pengesanan penyakit tanaman telah dibangunkan dan dinilai menggunakan set data penanda aras seperti PlantVillage, Cassava dan Tomato Leaf. Model yang dicadangkan, dengan hanya 23.04 juta parameter, mencatat peningkatan ketepatan sebanyak 0.3%, 0.8% dan 0.4% berbanding kaedah semasa, sekali gus mencapai keseimbangan yang baik antara prestasi dan kekompakan. Bagi pengecaman penyakit berbutir halus, kajian

ini turut memperkenalkan mekanisme perhatian hierarki. Pada tahap peta ciri, patch yang tidak penting dan tidak berkaitan disaring secara dinamik, dan ciri-ciri halus diekstrak hanya dari kawasan yang berkaitan rapat. Model ultra-ringan yang dihasilkan ini hanya mengandungi 1.3 juta parameter, namun mengatasi model arus perdana sekitar 0.3% dalam senario melibatkan sampel kecil dan penyakit yang hampir serupa, membuktikan keberkesanannya dalam mengenal pasti simptom awal dan kabur. Dalam aspek penyulingan pengetahuan, tesis ini mencadangkan mekanisme penyulingan tetingkap beralih yang peka terhadap sasaran. Kaedah ini mengira korelasi antara peta ciri guru dan pelajar, dan memperuntukkan berat penyulingan secara dinamik. Peta ciri yang besar dibahagikan kepada patch dan diproses menggunakan perhatian mengikut kumpulan serta strategi tetingkap beralih, bagi mengekalkan maklumat sempadan dan mengurangkan pengiraan berlebihan, seterusnya meningkatkan kecekapan pemindahan pengetahuan merentasi lapisan dan skala. Keputusan eksperimen menunjukkan bahawa bagi tugasan pengelasan ImageNet-1K dan segmentasi Pascal VOC, kaedah yang dicadangkan mencatatkan peningkatan purata ketepatan sekitar 1% berbanding pendekatan penyulingan sedia ada. Secara keseluruhannya, kajian ini membangunkan model Transformer yang cekap dan ringan untuk pengesanan penyakit tanaman, dengan peningkatan ketara dari segi ketepatan dan kelajuan. Ia juga mencadangkan mekanisme penyulingan baharu yang lebih mudah suai dan peka terhadap sempadan, sekali gus memperluas aplikasi Transformer dalam pemampatan model dan pembelajaran pemindahan. Penyelidikan ini menyediakan sumbangan teori dan panduan praktikal untuk sistem penderiaan pintar pertanian serta reka bentuk model penglihatan komputer yang ringan.

IMPROVED CROP DISEASE DETECTION USING HYBRID VISION TRANSFORMERS AND KNOWLEDGE DISTILLATION

ABSTRACT

In the field of computer vision, transformer architectures have emerged as powerful alternatives to traditional convolutional neural networks, offering superior global modeling capabilities and remarkable performance in various visual recognition tasks. However, their large parameter size and reliance on extensive training data, coupled with a lack of efficient structures for fine-grained feature extraction, hinder their deployment in precise crop disease detection scenarios. In the domain of knowledge distillation, conventional teacher–student frameworks often suffer from mismatches between intermediate feature maps due to architectural depth differences. Moreover, the high dimensionality of intermediate feature maps results in quadratic growth in computational cost when directly performing feature map-level distillation, significantly reducing efficiency. To address these challenges, this work first improves the multi-head self-attention mechanism in ViT by introducing a Group-wise Attention design. The attention computation is partitioned into multiple sub-groups, within which parameters are shared and local attention is computed independently. The outputs are then aggregated to retain global modeling capacity while reducing parameter count and computational overhead. Based on this mechanism, a lightweight ViT model for crop disease detection is constructed and evaluated on benchmark datasets such as PlantVillage, Cassava, and Tomato Leaf. The proposed model, with only 23.04M parameters, achieves accuracy improvements of 0.3%, 0.8%, and 0.4% over state-of-the-art methods, striking a favorable balance between performance and model compactness. For fine-grained disease recognition, a hierarchical attention mechanism is further introduced. At the feature map level, non-essential patches unrelated to disease are dynamically filtered out, and fine-grained features are extracted only from highly relevant regions. The resulting ultra-lightweight model contains merely 1.3M

parameters, yet outperforms current mainstream models by approximately 0.3% in scenarios involving small-sample and visually similar diseases, demonstrating strong capabilities in detecting early-stage, ambiguous, and similar disease symptoms. In the context of knowledge distillation, this thesis incorporates a Target-aware Transformer and proposes a Shifted Window Target-aware Distillation mechanism. By computing the correlation between intermediate feature maps of teacher and student networks, this method dynamically allocates distillation weights. The large feature maps are divided into patches and processed with group-wise attention and a sliding-window strategy to preserve boundary information and reduce redundant computation, thereby improving cross-layer and multi-scale knowledge transfer efficiency. Experimental results show that on both the ImageNet-1K classification and Pascal VOC segmentation tasks, the proposed method achieves an average accuracy improvement of around 1% over existing leading distillation approaches. In summary, this study constructs an efficient and lightweight Transformer-based model for crop disease detection, significantly enhancing recognition accuracy and inference speed. It also proposes a novel distillation mechanism with structural adaptability and boundary-awareness, expanding the applicability of Transformers in model compression and transfer learning. The research provides both theoretical insights and practical guidance for intelligent agricultural sensing systems and lightweight vision model design.

CHAPTER 1

INTRODUCTION

1.1 Overview

Artificial intelligence (AI) has rapidly advanced in recent years, offering transformative solutions across multiple domains. Among the various branches of AI, machine learning has played a pivotal role in enabling systems to learn from data and make intelligent decisions. Deep learning (LeCun et al., 2015), a subset of machine learning, has further elevated this capability by allowing models to extract complex features from vast datasets automatically. Building on these developments, computer vision has emerged as a key application of deep learning, enabling machines to interpret and analyze visual information accurately.

Computer vision has seen remarkable growth, beginning with the development of Convolutional Neural Networks (CNNs) (Krizhevsky et al., 2012) in the late 1990s. CNNs, popularized by models like AlexNet and ResNet, introduced the concept of hierarchical feature extraction using convolutional layers to analyze image data, revolutionizing tasks like image classification and object detection. However, CNNs primarily rely on localized receptive fields, which can limit their ability to capture long-range dependencies in images. To address this, the Vision Transformer (ViT) (Dosovitskiy et al., 2020) emerged, leveraging self-attention mechanisms (Vaswani et al., 2017) to capture global dependencies in images more effectively. Unlike CNNs, which process images hierarchically, ViT processes the entire image at once, allowing for more efficient global feature extraction. This transition from CNNs to ViTs marks a significant leap in the field of computer vision. In the context of smart agriculture, this advancement plays a crucial role in enhancing crop disease detection, as ViTs enable more precise and efficient identification of diseases across diverse crop species.

In the domain of crop disease detection, ViT have demonstrated significant potential. By capturing global features and contextual information from crop images,

Transformers can enhance the accuracy of disease detection and classification. This is particularly valuable for intelligent agriculture, where precise and reliable detection of crop diseases is critical for effective management and yield optimization. Furthermore, in fine-grained crop disease detection, Transformers demonstrate exceptional performance by integrating detailed feature representations across multiple scales and layers. This ability to capture intricate patterns and subtle variations in disease symptoms contributes to improved detection accuracy and finer classification.

The application of Transformers extends to knowledge distillation (KD) (Gou et al., 2021), where they play a crucial role in transferring knowledge from larger, more complex models to smaller, more efficient ones. The attention mechanism in Transformers guides the student network to focus more on the critical regions of the teacher network's feature maps during the intermediate layer knowledge distillation process in neural networks, enhancing the student's ability to generalize and perform well on various tasks with limited resources.

Recent advancements include the development of hierarchical and window-based approaches, such as the Swin Transformer (Yao & Shao, 2024), which address limitations in fine-grained feature extraction and model parameter efficiency. These innovations, along with techniques for parameter pruning, quantization, and hybrid models combining Transformers and CNNs, contribute to optimizing ViTs for practical applications. These developments aim to balance accuracy, computational cost, and memory usage, expanding the applicability of Transformers across diverse computer vision tasks and resource-constrained environments

This research focuses on designing and optimizing efficient ViT model architectures, exploring their applications in crop disease classification and fine-grained detection, and investigating their role in knowledge distillation to enhance practical effectiveness in various domains.

1.2 Background

Crop diseases pose significant threats to global food security by causing substantial yield losses and economic damage. Early and accurate detection of plant diseases is essential for timely intervention, limiting disease spread, and promoting sustainable agricultural production. Recently, advances in computer vision (CV) and deep learning have driven significant progress in automatic crop disease detection, surpassing traditional manual or rule-based methods.

CNNs have been widely adopted in agricultural image analysis due to their strong feature extraction capabilities. However, CNNs mainly focus on local receptive fields and layered convolutions (LeCun et al., 1998; He et al., 2016; Krizhevsky et al., 2017), which limits their ability to capture global contextual relationships, particularly in complex agricultural environments. To overcome this limitation, ViTs (Dosovitskiy et al., 2020) have been introduced, leveraging self-attention mechanisms (Vaswani et al., 2017) to model long-range dependencies. ViTs have achieved notable success across various vision tasks, including plant disease classification. Despite their effectiveness, ViTs typically require large-scale annotated datasets and possess high computational demands, challenging their deployment in resource-constrained scenarios.

In addition, crop disease detection under field conditions faces challenges such as subtle symptom differences, visual similarity among disease categories, varying lighting, occlusion, and background noise. These factors complicate fine-grained classification, demanding models with enhanced representational power.

Parallel to these developments, knowledge distillation has emerged as a critical technique in model compression and lightweight design. By transferring knowledge from a large, complex teacher model to a smaller student model, knowledge distillation enables substantial reduction in model size and computational cost while retaining performance. Among various distillation methods, effective transfer of intermediate layer knowledge has proven to be a vital factor for improving distillation quality and

enhancing the student's representation learning. This aspect is increasingly recognized as a key direction in current research on model lightweighting. Figure 1.1 shows the key milestones in the development of image classification neural networks.

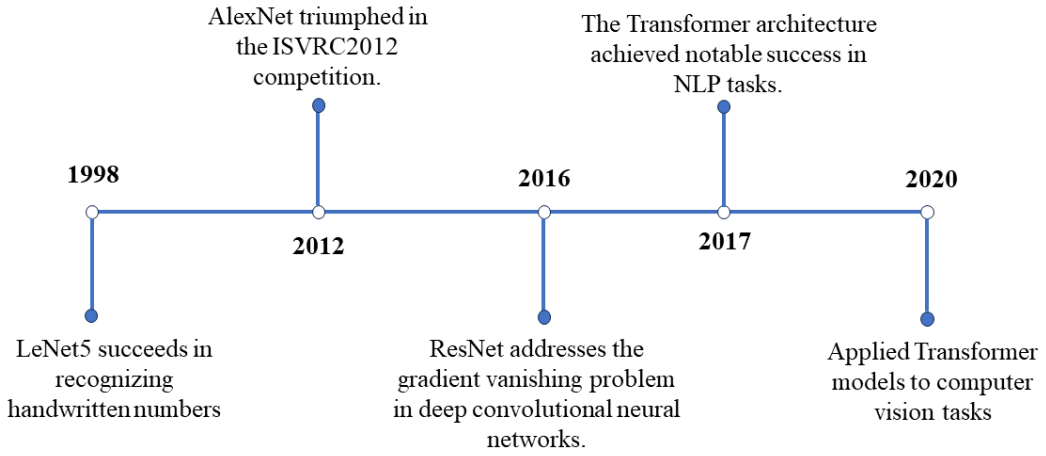

Figure 1.1: Key Milestones in the Development of Image Classification Networks

In various fields of computer vision, Transformers have demonstrated remarkable advantages. Their ability to capture long-range dependencies and integrate multi-scale features has led to significant improvements in performance. For instance, in medical imaging, Transformers enhance diagnostic accuracy by effectively analyzing complex visual patterns (Manzari et al., 2023), while in autonomous driving, they improve object detection and scene understanding (Prakash et al., 2021). Transformers have also been applied to the field of intelligent agriculture, where they demonstrate immense potential in crop disease classification. By effectively integrating multi-scale feature representations, Transformers significantly enhance the accuracy of disease detection. For fine-grained detection tasks, their ability to fuse multi-level information allows for precise differentiation between visually similar disease classes. Furthermore, in knowledge distillation tasks, the attention mechanism of Transformers efficiently guides the student network to focus on critical regions of the teacher network's feature maps, thereby improving the overall efficiency of knowledge transfer.

1.3 Research Motivation

The increasing global demand for food security necessitates the development of advanced technologies to ensure efficient and timely crop disease management. Early detection and accurate classification of crop diseases are crucial for minimizing agricultural losses, maximizing yields, and promoting sustainable farming practices. However, traditional machine learning and computer vision methods often struggle to efficiently address the complexities of crop disease detection. Moreover, the real-time detection of crop diseases imposes stringent requirements on model efficiency and lightweight design.

The task of crop disease detection involves analyzing leaf images, where the complexity of agricultural production environments further increases the difficulty of classification. Crop disease detection models not only need to focus on local disease features in the images but also must be capable of capturing long-range dependencies. By incorporating contextual information to achieve global perception, these models can more accurately understand the morphological characteristics of diseases, thereby enhancing detection performance. Additionally, early disease lesions are typically small, and the differences between various disease types may be very subtle, making fine-grained disease detection a popular research topic in this field.

The Transformer architecture is highly suited to address the aforementioned challenges. Its self-attention mechanism effectively captures long-range dependencies, enabling the model to focus on both local disease features and global context simultaneously. This capability enhances the model's ability to comprehensively understand the morphological characteristics of crop diseases. Furthermore, by applying Transformers to fine-grained disease detection, the model can handle subtle differences between early disease lesions and various disease types. These advantages make Transformers an ideal choice for improving the performance of crop disease detection models.

Moreover, lightweight model design is an important research focus in deep neural networks. The current classical knowledge distillation algorithms aim to transfer the knowledge learned by the teacher network to a smaller student network. In this transfer process, the student network needs to effectively focus on and inherit the key knowledge from the teacher network. This mechanism led us to draw a parallel with the attention mechanism in Transformers. Therefore, we attempt to apply Transformers to knowledge distillation to further explore the potential for model lightweight.

In summary, exploring and optimizing the application of Transformers in crop disease classification, fine-grained detection, and knowledge distillation not only advances the development of intelligent agriculture but also improves the accuracy and efficiency of crop disease detection, providing robust technical support for agricultural production. These research directions hold significant academic value and practical significance, laying a foundation for the broad application of computer vision technologies in the agricultural sector.

1.4 Research Problem

ViTs have shown strong potential in crop disease detection by effectively capturing long-range dependencies, outperforming traditional CNNs. Their token-based representations also offer structural advantages for aligning intermediate features, making them well-suited for knowledge distillation. However, ViTs are computationally intensive and heavily reliant on large annotated datasets, which limits their deployment in real-world agricultural settings, particularly on edge devices. Furthermore, while knowledge distillation is a promising solution for model compression, its application within Transformer-based architectures—especially at the intermediate feature level—remains underexplored. In response to these challenges, this study addresses the following research problems:

1. Although ViTs overcome CNNs' limitations in modeling global contextual information, they come with significant computational costs and large parameter

sizes. These factors increase the demand for extensive training data and hinder real-time deployment in agricultural environments with limited resources. It is therefore necessary to design lightweight ViT architectures that retain high detection accuracy while reducing model complexity and data dependency.

2. Beyond general classification, real-world agricultural environments present more complex challenges for crop disease detection. In particular, early-stage disease symptoms are often visually subtle and ambiguous, and many diseases exhibit high inter-class similarity. These factors complicate classification tasks and necessitate fine-grained recognition capabilities. Existing models often struggle to extract and discriminate subtle visual differences, especially under field conditions such as inconsistent lighting and background noise.

3. Knowledge distillation offers a promising approach to compress large models into lightweight student networks. However, traditional distillation methods suffer from feature misalignment due to inconsistent feature map sizes between teacher and student networks. This spatial mismatch limits the effectiveness of intermediate layer knowledge transfer. Transformer-based architectures, with their uniform token representation, naturally alleviate this mismatch and provide a more compatible framework for distillation. Despite this advantage, the use of Transformer-based knowledge distillation to reduce model's size remains largely unexplored.

1.5 Research Objectives

The primary objective of this research is to develop and optimize Vision Transformer architectures for efficient and accurate crop disease detection, with a particular emphasis on fine-grained classification. In terms of lightweight model design, the attention mechanism in Transformers can play a crucial role in the process of knowledge distillation. To achieve these goals, the research objectives are outlined as follows:

1. To analyze the limitations of standard Vision Transformers in agricultural scenarios with limited computational resources. Propose a novel hybrid Transformer-based model that incorporates advanced group-wise techniques. This model aims to maintain high detection accuracy while significantly improving computational efficiency, enabling real-time deployment in resource-constrained agricultural environments.

2. To address the critical need for fine-grained crop disease detection, which requires distinguishing visually similar symptoms across different disease types or growth stages, this study first identifies the limitations of general classification models in capturing subtle inter-class differences. Building on this, we propose a Coarse-to-Fine Attention Network (C2FAN) that hierarchically refines multi-scale visual features to bridge this gap, enabling more accurate and interpretable disease diagnosis.

3. To investigate the limitations of traditional knowledge distillation methods caused by spatial feature misalignment. By leveraging the uniform token representation of Transformer architectures, the research aims to facilitate more accurate and efficient intermediate knowledge transfer, thereby enabling the construction of compact student models with high performance. This study seeks to bridge the gap in current research by applying Transformer-based distillation strategies to significantly reduce model size while preserving accuracy.

1.6 Research Scope

The scope of this research is centered on improving the performance and efficiency of Transformer-based models for crop disease detection, using diverse datasets to ensure comprehensive evaluation. This study focuses on addressing computational challenges and optimizing the models for practical applications in agriculture. It also emphasizes fine-grained analysis of crop disease images, aiming to enhance detection accuracy by progressively refining the focus from coarse to detailed features. The research also seeks

to improve resource efficiency in deep neural networks, ensuring high performance is maintained while reducing computational overhead.

To ensure a robust evaluation, the research utilizes several datasets. These include publicly available crop disease detection datasets, similar datasets designed for fine-grained disease detection among similar crop diseases, and standard datasets for testing classification performance. The public benchmark dataset is employed for semantic segmentation tasks, ensuring that the models are evaluated against widely recognized standards in the field.

Overall, this research aims to enhance both the effectiveness and computational efficiency of Transformer-based models for crop disease detection, with a strong focus on real-world applicability in agricultural contexts. Additionally, the study leverages the strengths of the attention mechanism in Transformers, successfully applying it to knowledge distillation, further improving the resource efficiency and performance of deep neural networks.

1.7 Research Contribution

The efficiency of neural networks is a critical metric for model evaluation, reflecting both performance and computational complexity. To address these challenges, we have optimized the self-attention mechanism within the Vision Transformer (ViT) framework and proposed the EfficientNet Convolutional Group-Wise Transformer (EGWT) . This enhanced model architecture is specifically applied to crop disease detection within intelligent agriculture contexts. Our work assesses the model's stability and resource requirements in practical applications, ensuring efficient operation of neural network-based detection systems in real-world agricultural production environments.

Furthermore, leveraging the advantages of Transformer models, we propose an innovative method for knowledge distillation from intermediate layers. This method enhances the Target-aware Transformer (TaT) model (Lin et al., 2022) by enabling

the model to utilize attention mechanisms to transfer knowledge from the intermediate feature layers of a large neural network to those of a smaller network. The integration of the Shifted Window (Swin) design facilitates the student network's ability to learn richer features from the teacher network. This approach significantly improves performance in image classification and semantic segmentation tasks, building upon the foundational capabilities of the Target-aware Transformer.

Our contributions advance the application of Transformer models in crop disease classification, fine-grained detection, and knowledge distillation, providing new insights into model efficiency, performance optimization, and practical deployment in agricultural settings.

1.8 Thesis Organization

This thesis is organized as follows:

Chapter 1 introduces the research overview, and background, defines the research problem, outlines the objectives and contributions, and provides an overview of the thesis structure.

Chapter 2, we review the existing literature on ViT and related studies, critically analyze previous works, and identify research gaps that this thesis aims to address.

Chapter 3 presents the methodology adopted in this research, detailing the design of the proposed model, including the modifications made to improve crop disease detection accuracy, and describes the experimental setup, dataset preparation, and evaluation metrics used.

Chapter 4 discusses the experimental results, providing an in-depth analysis of the model's performance, comparing it to baseline methods, and presenting a thorough evaluation using metrics such as accuracy, precision, recall, and computational efficiency.

Chapter 5 concludes the thesis by summarizing the findings, discussing the contributions to the field, and suggesting directions for future research based on the limitations and insights gained from this study.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter presents the current research progress of Transformer architectures in crop disease detection tasks. With the continuous advancement of deep learning technologies, the application of Vision Transformer (ViT) and its variants in smart agriculture has garnered increasing attention, particularly in crop disease detection, fine-grained detection, and knowledge distillation. Section 2.2 provides an overview of ViT, detailing its processing flow in image classification tasks. Section 2.3 focuses on reviewing the application of ViT in crop disease detection, while Section 2.4 further discusses its existing application in fine-grained detection. Section 2.5 reviews the Transformer architectures as a lightweight model, particularly in knowledge distillation. Finally, Section 2.6 summarizes the chapter.

2.2 Overview of Vision Transformer

The Transformer was initially proposed for the field of NLP and achieved great success in that domain. Inspired by this success, Dosovitskiy et al. (2020) seeks to validate the Transformer architecture's effectiveness in computer vision(CV). By converting image data into a sequence format similar to text data, Transformer models can effectively capture and process the structured information in images. This approach enables Transformers to not only be applied in natural language processing but also to handle complex visual tasks, leveraging their sequence modeling capabilities to extract feature patterns from images. Figure 2.1 presents the framework of the Vision Transformer (ViT) model. In brief, the model consists of two main modules:

1. Linear Projection of Flattened Patches (Embedding Layer).
2. Transformer Encoder (The detailed structure is provided in Figure 2.3).

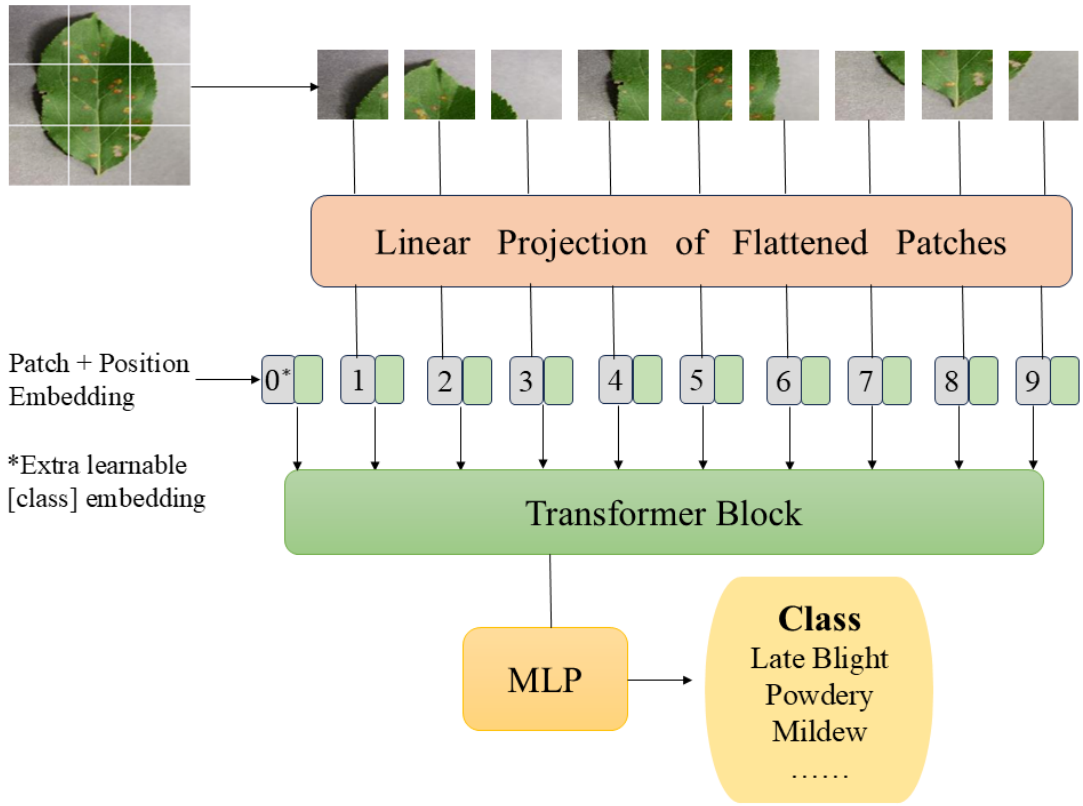

Figure 2.1: The Framework of Vision Transformer (ViT) Model

2.2.1 Patch Embedding Layer

The original images in the dataset, with assumed dimensions of $H \times W$ (where H and W denote the height and width of the image, respectively), are divided into several patches, each of size $P \times P$. These two-dimensional patches (ignoring the channel dimension) are flattened into one-dimensional vectors. The flattened patches are transformed into input tokens through a linear projection layer, after which the class token and positional encoding are added to create the final input for the transformer encoder. The class token is input into the network along with the preceding feature sequences for feature extraction and serves as a learnable parameter that incorporates class information. This class token is represented as 0^* in Figure 2.1, which increases the length of the flattened input vector by one.

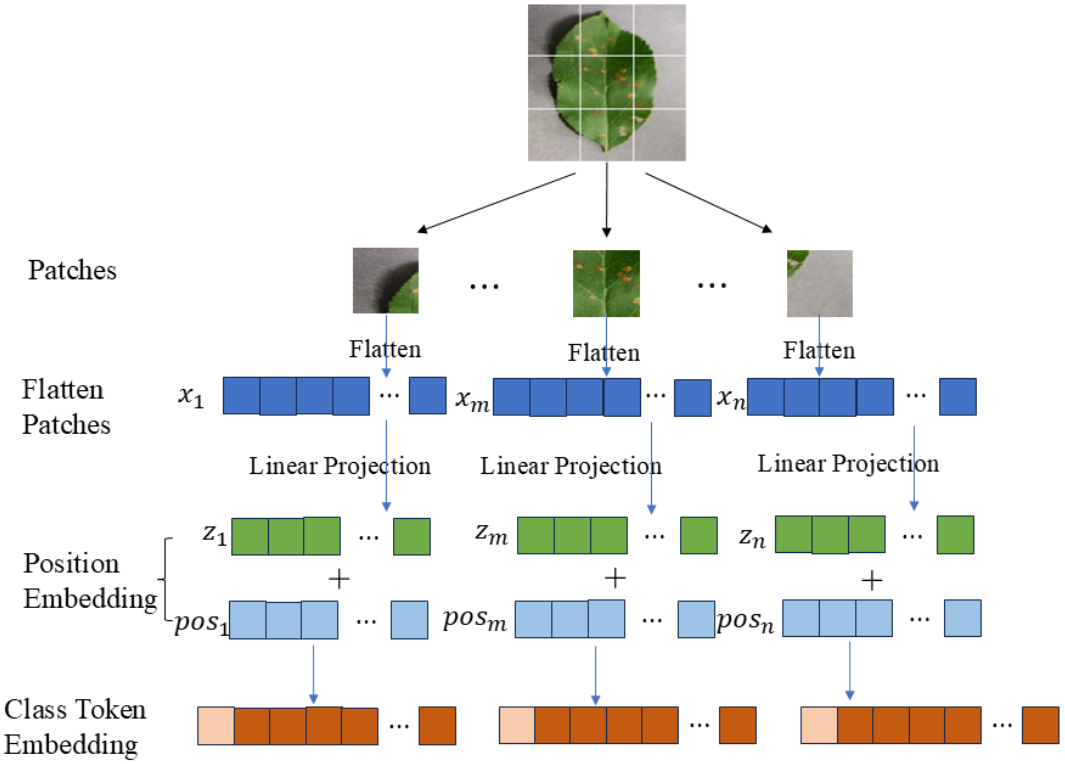

Figure 2.2: The Patch Embedding Layer.

As shown in Figure 2.2, the image is divided into N patches of size $P \times P \times C$. The calculation of N is as follows:

$$N = \frac{H \times W}{P \times P} \tag{2.1}$$

Each patch is flattened into a vector of length P^2, and then a linear transformation (typically represented by a weight matrix $\mathbf{E} \in \mathbb{R}^{(P^2 \times C) \times D}$, where D is the embedding dimension of the model) is applied to map this vector into a D-dimensional space. For each patch, this process can be represented as:

$$z_i = \mathbf{E} \cdot \text{flatten}(x_i) + b \tag{2.2}$$

where x_i is the input of the i-th patch, z_i is the corresponding embedded vector, b is the bias.

Due to the sequential nature of the input, it is necessary to add positional information to each input vector. In the ViT model, the method used is identical to the technique

commonly employed in NLP tasks, namely positional encoding (Chu et al., 2021), which enhances the model's ability to perceive the order of the sequence. This approach ensures that the model, when processing visual data, can capture the implicit positional information in the input, similar to how it handles textual sequences, thereby improving its understanding of the global structure. The positional embedding be denoted as:

$$z_i^{pos} = z_i + pos_i \tag{2.3}$$

where pos_i is the positional vector, and z_i^{pos} are the vectors of each patch after adding positional embeddings.

A specific class token is added at the beginning of the input sequence to enable the network to learn a global representation at the image level. The embedding vector for this class token is denoted as c.

$$z_0 = c \tag{2.4}$$

In ViT, the input sequence to the Transformer is given by:

$$Z = [z_0; z_1^{pos}; z_2^{pos}; \cdots ; z_N^{pos}] \tag{2.5}$$

where z_0 is the class token.

2.2.2 Transformer Encoder

Similar to Convolutional Neural Networks (CNNs), a backbone is needed for feature extraction. In ViT, the Multi-Head Attention (MHA) mechanism is the key component for feature extraction (see Figure 2.3). The sequence Z given by Equation 2.5 is fed into the encoder for feature extraction.

2.2.2(a) Layer Normalization

In the Figure 2.3, the input image is initially processed through a Layer Normalization (LN) (Ba et al., 2016), which stabilizes the training process and, in turn, enhances the model's convergence rate and overall performance. Given an input sequence

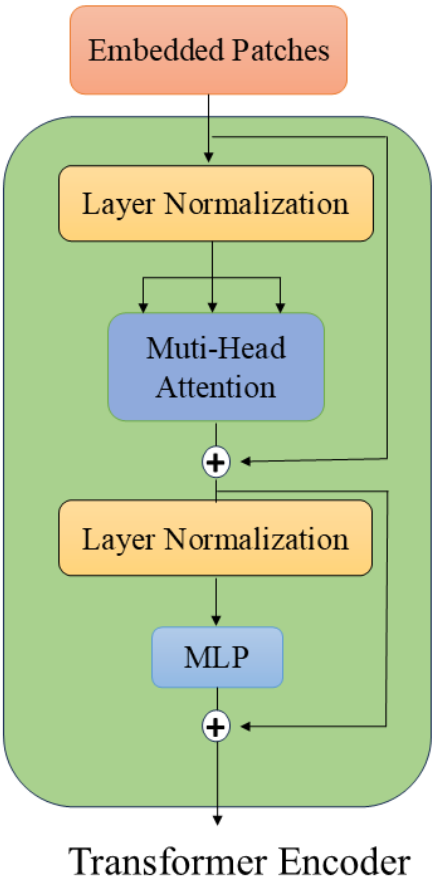

Figure 2.3: Flowchart of the Transformer Encoder Framework.

$Z = z_0; z_1^{pos}; z_2^{pos}; \cdots; z_N^{pos}$, the mathematical formulation is as follows:

$$\hat{z}_i = \frac{z_i - \mu}{\sqrt{\sigma^2 + \epsilon}} \tag{2.6}$$

where, μ represents the mean, σ^2 denotes the variance, and ϵ is a small constant used to prevent division by zero. Normalizing the vector z_i gives:

$$\mathrm{LN}(z_i) = \gamma \hat{z}_i + \beta \tag{2.7}$$

where γ and β are learnable parameters for scaling and shifting, respectively. The final output of Layer Normalization is:

$$\mathrm{LN}(Z) = \{\mathrm{LN}(z_1), \mathrm{LN}(z_2), \ldots, \mathrm{LN}(z_n)\} \tag{2.8}$$

This normalization is performed at each position in the sequence, helping to stabilize the training process and improve the model's performance.

2.2.2(b) Multi-Head Attention

Multi-Head Attention is an extended self-attention(SA) mechanism (Niu et al., 2021) that divides the input data into multiple subspaces, or "heads," and performs attention calculations independently for each subspace. Each head captures different features or relationships, and these features are subsequently combined to achieve a richer representation. The constituent component of multi-head attention is self-attention. Self-attention is a crucial mechanism for modeling the relationships between image patches. It enables the model to weigh each element in the input sequence, capturing dependencies between elements and thereby generating more comprehensive feature representations. For an input sequence $\boldsymbol{Z} = [z_1, z_2, \cdots, z_n]$, where each z_i is a vector of length d and the total sequence length is n, each input vector is first linearly mapped into three components: Query (q), Key (k), and Value (v). Self-attention is then computed via dot product. The self-attention formulas are expressed as follows:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right) V \tag{2.9}$$

where: $\mathbf{Q} = \mathbf{Z}\mathbf{W}^q$, $\mathbf{K} = \mathbf{Z}\mathbf{W}^k$, $\mathbf{V} = \mathbf{Z}\mathbf{W}^v$, d_k is the dimension of the key vectors, $\mathbf{W}^q$, $\mathbf{W}^k$, and $\mathbf{W}^v$ are the weight matrices used to generate queries, keys, and values, respectively. The attention calculation for a sequence of length 2 is shown in Figure 2.4.

After the attention is computed for each head, the outputs of all heads are concatenated. Specifically, each head i computes its attention as:

$$\text{head}_i = \text{Attention}(Q_i, K_i, V_i) \tag{2.10}$$

. The outputs from all heads are then concatenated along the feature dimension:

$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\text{head}_1, \text{head}_2, \ldots, \text{head}_h), \tag{2.11}$$

where h is the number of heads. Finally, the concatenated output is projected back into the original dimension via a linear transformation with a weight matrix W^O, resulting

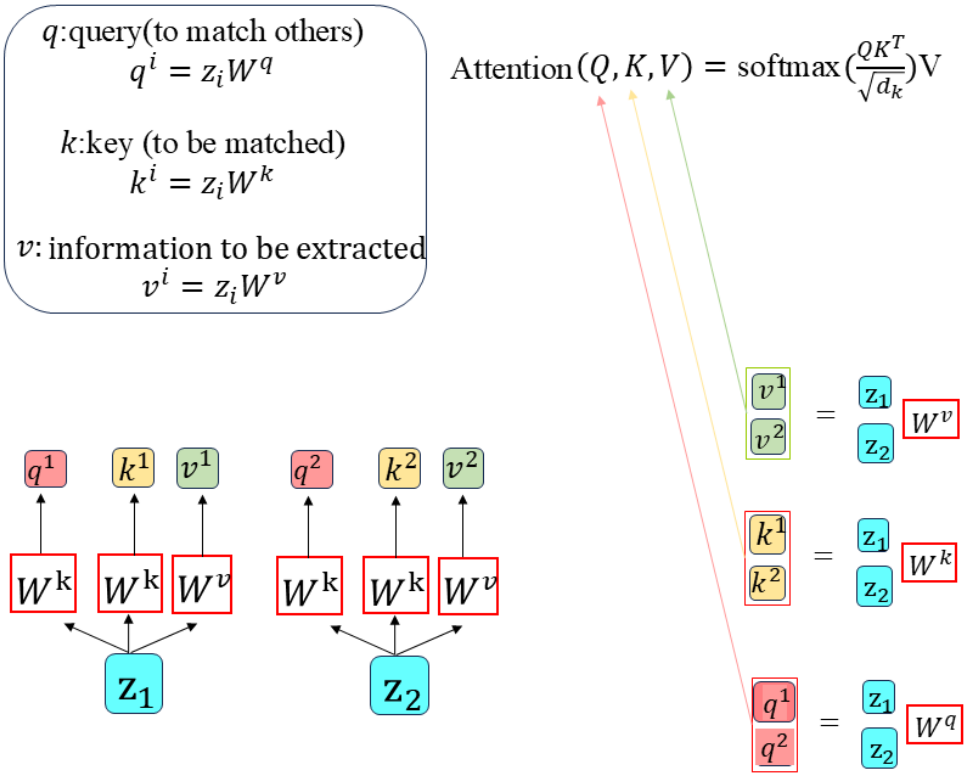

Figure 2.4: Illustration of Self-Attention calculation.

in the final output:

$$\text{MultiHead}(\mathbf{Z}) = \text{Concat}(\text{head}_1, \text{head}_2, \ldots, \text{head}_h)\mathbf{W}^O \tag{2.12}$$

Subsequently, the output is added to the residual connection, followed by normalization:

$$\mathbf{Z}_{\text{att}} = \text{LayerNorm}(\mathbf{Z} + \text{MultiHead}(\mathbf{Z})) \tag{2.13}$$

This allows the model to capture different aspects of the input through multiple attention heads, with each head focusing on different parts of the input, thereby enhancing the model's ability to learn complex patterns and relationships.

2.2.2(c) Multi-Layer Perceptron

In the ViT architecture, the Multi-Layer Perceptron (MLP) (Taud & Mas, 2018) is employed for feature processing and transformation following the application of

the self-attention mechanism. MLP is a feedforward neural network (Svozil et al., 1997; Fine, 2006) comprising multiple layers of fully connected (linear) neurons and activation functions (Rasamoelina et al., 2020). The formula for MLP is given by:

$$\mathbf{Z}_w = \mathrm{FC}_2(\mathrm{ReLU}(\mathrm{FC}_1(\mathbf{Z}_{\mathrm{att}}))) \tag{2.14}$$

where

$$\mathrm{FC}_1(x) = xW_1 + b_1,$$

$$\mathrm{ReLU}(x) = \max(0, x),$$

$$\mathrm{FC}_2(x) = xW_2 + b_2$$

Here, W_1 and W_2 are learnable weight matrices, and b_1 and b_2 are bias vectors. The ReLU activation function introduces non-linearity into the network.

The output layer transforms the activation results of the hidden layer into the final prediction output through the softmax activation function, yielding the predicted probability for each class. The formula is as follows:

$$\mathrm{Softmax}(z_w) = \frac{e^{z_w}}{\sum_{j=1}^{n} e^{z_w}} \tag{2.15}$$

For an input vector, the softmax function converts each element z_w into a probability value that lies within the range $[0, 1]$.

2.2.3 Limitation of ViT

Despite the advantages of ViT in handling long sequence data, its structure has some inherent limitations. Firstly, while the global self-attention mechanism in ViT excels at extracting global features, it tends to overlook the importance of local information, particularly at the initial stages. This results in suboptimal performance when ViT is applied to tasks that require high-resolution fine-grained features or the recognition of local textures (Wang et al., 2021). Secondly, due to the absence of convolution operations found in CNNs, ViT's structure lacks support for spatial translation invariance, which may reduce its robustness to small-scale transformations and noise (Liu et al.,

2021b). Additionally, the structure of ViT typically demands a significant number of parameters and computational resources to sustain the computation of its multi-layer self-attention mechanism, making it less suitable for resource-constrained environments. Furthermore, ViT has limited capacity to handle multi-scale information; its original structure was not designed to integrate multi-scale features present in natural images, leading to subpar performance in complex scenes or fine-grained tasks when compared to CNNs with multi-scale feature fusion or modified ViT variants (Xie et al., 2021). To address these structural limitations, researchers often introduce multi-scale feature extraction modules, hybrid convolution, and self-attention mechanisms, or other architectural enhancements (Yuan et al., 2021).

2.3 Vision Transformer in Crop Disease Detection

Crop disease detection holds significant importance in modern agriculture, serving as a critical component of precision agriculture. Early and accurate identification of plant diseases not only helps improve crop yield and quality but also reduces the excessive use of pesticides, thereby mitigating the environmental impact of agricultural production. Traditional methods rely on manual observation and expert judgment, which are labor-intensive, time-consuming, and prone to human error, thus limiting detection efficiency and accuracy. With the rapid advancement of deep learning technology, automated disease detection has become increasingly feasible. Particularly, models based on CNNs and ViTs demonstrate strong capabilities in image recognition and feature extraction, showing promise for significantly enhancing the efficiency and accuracy of crop disease detection and providing more intelligent solutions for agricultural production.

CNNs have long been the cornerstone of image-based crop disease detection, with numerous studies, such as those by Saeed et al. (2021) and Kurmi et al. (2022), demonstrating that CNNs can achieve high accuracy in identifying various crop diseases. However, CNNs extract feature information through local receptive fields (Qin et al.,

2018). In plant disease image detection, shallow CNNs primarily capture low-level features like edges, textures, and simple shapes, while deeper layers are required to discern more complex structural information. This layer-by-layer approach to feature extraction, however, presents limitations in fine-grained classification tasks and complex scenarios. ViTs address this limitation by utilizing self-attention mechanisms to establish long-range dependencies across different regions of an image, thereby directly capturing global contextual information without relying on CNNs' progressive feature aggregation. The ViT architecture not only surpasses the limitations of local receptive fields but also efficiently extracts rich multi-scale and global features from input images, demonstrating superior performance in image feature extraction tasks.

Over the past decade, transformer-based models and transfer learning have gained considerable traction, demonstrating significant potential in crop disease identification, as highlighted in studies such as Thakur et al. (2021) and Selvi et al. (2024). In the domain of crop leaf image recognition, models like ViT-SmartAgri (Barman et al., 2024), a smartphone-based Vision Transformer (ViT) developed for tomato disease detection, have shown promise. Trained on a dataset of 10,010 images, it surpasses traditional models like Inception V3 (Szegedy et al., 2015), achieving higher accuracy. This underscores ViT's potential in smart agriculture, enhancing crop disease detection through its self-attention mechanism, which captures spatial relationships and global context, improving precision in disease recognition.

GreenViT (Parez et al., 2023) explores the use of Vision Transformers (ViTs) for plant disease detection to enhance precision agriculture. The model's performance was evaluated against state-of-the-art CNN models using standardized datasets, demonstrating its superior capability and effectiveness compared to conventional approaches. Similarly, LeafViT (Keerthan Bhat et al., 2022), also based on the ViT architecture, utilizes the self-attention mechanism for in-depth plant leaf disease detection. Experimental results indicate that LeafViT excels in maintaining high accuracy, even under high-resolution and complex background conditions, enhancing crop health monitor-

ing's reliability and efficiency in real-world applications, and proving its potential for practical deployment.

These ViT-based approaches have demonstrated effectiveness in crop disease detection but also exhibit limitations. Firstly, ViT relies heavily on large-scale datasets for training, which are often scarce in the agricultural sector. When processing high-resolution images, ViT's computational complexity is high, increasing resource demands, and making it unsuitable for resource-constrained environments. Additionally, lacking convolutional operations, ViT may be less efficient than traditional CNNs in capturing local features and fine-grained disease patterns. Although ViT performs well in some tasks, its accuracy may lag behind hybrid models or CNNs when dealing with small datasets or complex scenarios, affecting stability and effectiveness. Therefore, further exploration of more efficient ViT models is needed to enhance performance and adaptability in diverse conditions for improved crop disease detection.

2.3.1 Base ViT Models in Crop Disease Detection

Due to its capacity for modeling long-range dependencies and capturing fine-grained features in image classification tasks (Dosovitskiy et al., 2020; Khan et al., 2022), the ViT has shown significant potential in plant disease detection. Traditional CNNs mainly rely on local convolution operations, which may struggle to detect complex structures and subtle differences in plant disease images (Boulent et al., 2019). By contrast, ViT's self-attention mechanism enables it to capture comprehensive global information, significantly improving detailed disease detection accuracy. In agriculture, plant diseases seriously threaten crop health and yield, ultimately affecting food security. Applying ViT in plant disease detection supports high-precision and fine-grained identification, facilitating early warning and targeted interventions for farmers. This approach mitigates the risk of disease spread and enhances crop quality and yield (Dhanya et al., 2022; Parez et al., 2023).

Early studies explored the replacement of traditional CNN structures with ViT architectures in the context of crop leaf disease detection. For instance, Thai et al. (2021) and Li et al. (2022) directly applied the ViT model to leaf disease detection tasks, achieving superior performance over CNNs, highlighting ViT's more comprehensive feature extraction capabilities. In particular, the application of a fine-tuned and pre-trained ViT model to rice disease detection (Rachman et al., 2024) demonstrated ViT's strength in handling large-scale data, showcasing its ability to effectively utilize extensive datasets and benefit from increased model capacity.

While ViT has shown promise in crop leaf disease recognition, challenges remain, such as limited data, high computational demands, inconsistent environmental conditions, reduced generalization, and accessibility barriers. To address these issues, advanced Transformer models have been developed. For example, the multi-scale contextual feature extraction network introduced by Salamai et al. (2023) leverages lesion-aware visual transformers to capture features across multiple scales and channels, improving both efficiency and generalizability compared to ViT on rice disease datasets. Additionally, Pacal (2024) advanced crop leaf disease recognition by replacing the CNN and MLP components within the ViT architecture.

Building on ViT, Liu et al. (2021b) proposed the Swin Transformer by incorporating a local window attention mechanism. In this design, self-attention is applied within each window, and cross-window information exchange is achieved via shifted window operations at various levels. This approach facilitates the progressive fusion of local and global information, positioning the Swin Transformer as a critical backbone model for computer vision tasks. Inspired by this innovation, a method utilizing step-wise small patch embeddings was developed to enhance feature extraction without increasing the model's parameter count. This approach, when applied to the Swin-Tiny model, has exhibited excellent performance in cucumber leaf disease recognition tasks (Wang et al., 2022).

The evolution of Vision Transformers (ViTs) has significantly advanced the capabilities of deep learning models, these enhancements aim to elevate model performance while addressing computational constraints, thereby increasing both the efficiency and practical applicability of Transformer models, such as the V2X-ViT (Xu et al., 2022), which emphasizes ViT's versatility and robustness, particularly in dynamic and noisy environments. Although primarily focused on 3D object detection within vehicular networks, this research has implications for crop disease detection, where environmental variability poses similar challenges. ViTs' adaptability to different domains underscores their potential for high accuracy and resilient performance in diverse conditions.

Similarly, the Slide-Transformer study demonstrates another innovative approach by incorporating local attention mechanisms, which substantially enhances computational efficiency without compromising model performance (Pan et al., 2023). This technique is particularly advantageous for processing large-scale agricultural datasets, where precise crop disease detection is critical.

In summary, while base ViT models have demonstrated promising potential in advancing crop disease detection, several critical challenges hinder their widespread adoption. Above existing models are trained and evaluated on well-curated datasets, which fail to reflect the complexities of real-world agricultural environments characterized by occlusion, background noise, and variable lighting. Additionally, the high computational cost of ViT architectures poses obstacles to their deployment on edge devices commonly used in rural areas. A further limitation lies in the lack of model interpretability, which undermines user trust and practical applicability. Moreover, the generalizability of ViTs across different crop types, disease stages, and geographical regions remains underexplored. The current dependence on large-scale labeled datasets highlights the urgent need for data-efficient learning strategies, such as self-supervised or few-shot learning. Therefore, to fully realize the potential of ViTs in smart agriculture, future research must address issues of generalization, scalability, and explainability by developing lightweight, robust, and interpretable models, validated