Vision Transformer Algorithm for Plant Disease Detection: A Systematic Literature Review

Reza Al Husna
Doctoral program, Information Technology, Universitas Gunadarma
Jakarta, Indonesia
rezaalhusna@staff.gunadarma.ac.id

Lintang Yuniar Banowosari
Department of Informatics Program Universitas Gunadarma
Jakarta, Indonesia
lintang@staff.gunadarma.ac.id

Tubagus Maulana Kusuma
Department of Technology and Engineering
Universitas Gunadarma
Jakarta, Indonesia
mkusuma@staff.gunadarma.ac.id

Tri Handhika
Department of Computational Mathematics Studies
Universitas Gunadarma
Jakarta, Indonesia
trihandika@staff.gunadarma.ac.id

2024 Ninth International Conference on Informatics and Computing (ICIC) | 979-8-3315-1760-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICIC64337.2024.10956258

***Abstract*—Advances in deep learning have significantly impacted various sectors, including agriculture, by improving efficiency and sustainability. Computer vision-based plant disease detection that utilizes deep learning models such as Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) has replaced manual methods that require specialized skills. Vision Transformers (ViTs) with self-attention mechanisms excel at recognizing complex patterns in images of diseased plants, enabling early detection and more effective management. This study conducted a systematic literature review to evaluate the use of feature fusion-based transformer algorithms in plant disease detection and to identify current research trends, challenges, and future opportunities. The results show that ViTs can improve accuracy and efficiency in plant disease diagnosis, particularly under diverse environmental conditions. However, the adoption of this technology faces obstacles such as the need for large annotated datasets, significant computational costs, and variability in field conditions that can affect model performance. This research highlights the need for further innovation to address these challenges and expand the accessibility and reliability of deep learning-based plant disease detection for the global agricultural sector.**

Keywords*—*deep learning, vision transformers, feature fusion, plant disease detection, self-attention mechanism

I. INTRODUCTION

In recent decades, advances in artificial intelligence (AI) technology, particularly in the application of deep learning, have made a significant impact in various fields, including agriculture. Convolutional Neural Network (CNN) algorithms have become a popular choice for image analysis due to their ability to efficiently extract spatial features from images [1]. CNNs enable the detection of various plant diseases with high accuracy based on visual patterns in leaves or [1]. For example, CNNs have been used in the classification of foliar diseases in tomato plants with a satisfactory success rate [2].

The adoption of deep learning in agriculture is increasing due to the need to address crop loss and increase productivity, especially in areas with minimal access to agricultural experts. Conventional methods for detecting plant diseases often rely on visual inspection by trained experts, which often takes a long time, produces inconsistent results, and is error-prone. In contrast, deep learning-based systems can analyze large amounts of image data in real-time, providing accurate and rapid diagnosis, which enables early intervention and more effective disease management.

Along with the development of technology, Transformers, which were originally developed for natural language processing [3], began to be applied in image analysis tasks. The Transformer architecture offers the ability to capture long-term relationships between features in an image through a self-attention mechanism, which allows the model to consider the entire image without losing important information [4]. The application of Transformers in agriculture is gaining attention with improved performance in plant disease classification lye[5].

A further development of the Transformer is the Vision Transformer (ViT), which is specifically designed for image processing tasks [6]. ViT transforms images into a sequence of patches and utilizes the Transformer architecture to perform image classification more effectively. This is particularly useful in detecting plant diseases that are difficult to recognize with the naked eye, such as viral or fungal infections that do not show obvious symptoms on the leaf surface [7]. The advantage of ViT over CNN lies in its ability to capture more detailed visual information without the need for pooling operations that can remove important information [8]. ViT has shown superior performance in detecting plant diseases in complex environments [9]. However, ViT tends to require more computational resources compared to traditional methods, which can be an obstacle, especially in agricultural areas with limited access to technology. In addition, ViT can also face difficulties in dealing with background complexity, which often obscures the target object in the image.

One approach to overcome the drawbacks of Vision Transformer is the application of feature fusion techniques. By combining information from various feature sources, such as texture, color, and shape, models can improve detection accuracy and reduce the risk of misclassification. Feature fusion can not only enrich the feature representation used by the Vision Transformer but also improve its ability to identify plant diseases that are difficult to detect. In this context, further research is needed to explore the effectiveness of feature fusion in improving the performance of Vision Transformer, especially in situations where datasets are limited and disease symptoms are not obvious.

In this study, we will conduct a Systematic Literature Review (SLR) on the application of feature fusion-based transformer algorithms in plant disease detection. The structure of this paper is as follows: Section 2 discusses the research methodology. Section 3 describes the results obtained. Finally, section 4 discusses the implications and contributions of this research.

II. Methodology

The main objective of a Systematic Literature Review (SLR) is to provide a comprehensive and structured review of the research that has been conducted in a particular field, by identifying, assessing, and synthesizing findings from relevant studies [10]. SLR serves to reveal current research trends, existing gaps in the literature, as well as future research opportunities, so as to provide a solid basis for further research and well-informed practice [11].

To achieve the objectives of this research, a systematic literature review (SLR) was conducted using VOS viewer to analyze and synthesize existing research to identify research trends and potential future research. Figure 1 shows the steps in a systematic literature review that includes literature search, study selection, and synthesis of findings. The final step in the process is to identify future research opportunities described in the implications and contributions section:

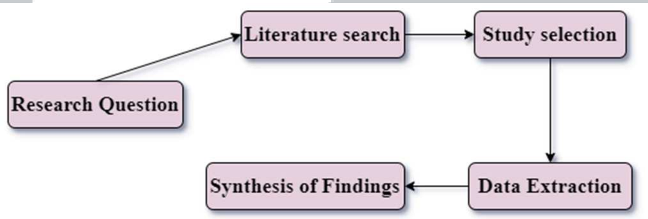

Fig. 1. Systematic Literature Review Steps.

A. Research Questions

Research questions play a crucial role in a systematic literature review (SLR) in the field of plant disease detection for several reasons, including helping to clearly define the research focus and scope, identifying key terms and concepts to be used in the search strategy [12], selecting relevant databases and literature sources, and facilitating the data acquisition process [13]. Table 2 presents the research questions formulated based on the Population, Intervention, Comparison, Outcome, and Context framework from Table 1. In this context, Population refers to the type of data or plant domain that is the subject of application of the disease detection method. Intervention refers to the specific detection technique or method used for plant disease identification. Comparison includes the evaluation of differences in the applied detection methods, such as the comparison of detection results with and without the application of a particular method. Results reflect the effectiveness of the detection method measured to assess the performance of the system in identifying plant diseases. Finally, context refers to the specific conditions or environment in which this research was conducted.

TABLE I. PICOC

	Contents
Population	Agricultural crops affected by plant diseases
Intervention	Deep learning models, feature fusion methods, CNN, Transformer
Comparison	Disease detection performance with vs. without specific feature fusion methods
Outcome	Detection accuracy, computational efficiency, robustness in varying environmental conditions
Context	Real-world agricultural diseasee detection scenarios

TABLE II. Research Questions

No.	Question
1.	RQ1: How do different feature fusion methods impact the performance of vision transformer models in terms of accuracy and computational efficiency?
2.	RQ2: What metrics are used to evaluate the effectiveness of different deep learning approaches in detecting plant disease?
3.	RQ3: What are the major challenges in applying feature fusion techniques with vision transformer models for accurate plant disease detection?

B. Literature Search

An in-depth literature search was conducted through leading academic databases, such as IEEE Xplore, ScienceDirect, MDPI, and Scientific reports. The keywords used in this search included “plant disease detection,” “deep learning,”, “transformer,” and “feature fusion.”. Table 3 shows detailed search queries including identify key concepts, generate keywords and synonyms.

TABLE III. Search Queries

Process	Results
Identify Key Concepts	{“plant disease”, “deep learning”, “transformer”, “feature fusion”}
Generate keywords and synonyms	• Plant disease detection = {leaf disease detection, crop disease identification, agricultural disease diagnosis, plant pathology recognition} • Deep learning = {neural network, convolutional neural network (CNN), AI-driven learning} • Transformer = {vision transformer, attention mechanism, self-attention, transformer network} • Feature fusion = {multi-modal fusion, feature combination, data integration, hybrid features}
Combine keywords with boolean operators	• ("plant disease” OR “leaf disease” OR “crop diseases” OR “agricultural disease” OR plant pathology recognition”) • AND (“deep learning” OR “neural network” OR convolutional neural network” OR “AI-driven lerning”) • AND (“transformer” OR “vision transformer” OR “attention mechanism” OR “self-attention” OR (“transformer network”) • AND (“feature fusion” OR multi-modal fusion” OR “feature combination” OR “data integration” OR “hybrid features”)

C. Study Selection

Relevant papers were selected based on certain criteria, including their relevance to the research topic, methodological quality, and validity of the study results. The selected research should include data on the application of artificial intelligence in the agricultural sector, particularly in the context of plant disease detection. Table 4 shows criteria which were used in this study.

TABLE IV. CRITERIA

Criteria type	Inclusion	Exclusion
Period	From 2024-2020	Prior 2020
Type of Literature	Research articles, conference, book chapters, review articles, government document	Report, newsletter, working papers
Type of source	Conferences or journals	Books
Language	Writen in English	Not writen in English
Accessibility	-	Not Accessible

D. *Data Extraction*

The data extraction process involves systematically collecting and organizing relevant data from the studies that have been included in the review. Key elements in data extraction include research characteristics, methodology, quality indicators, conclusions, recommendations, and strengths and weaknesses of the research. The main goal of the data extraction process is to ensure that the data captured is relevant to our research questions. The main objective of the data extraction process is to ensure that the extracted data fits into our research scheme.

E. *Synthesis of Findings*

Integration and analysis of findings from various studies were conducted to identify patterns, trends, as well as gaps in current research and future policy. This review evaluates the potential, challenges and benefits of applying deep learning in optimizing plant disease diagnosis.

III. RESULT AND DISCUSSION

A. *Result Findings*

Based on the parameters described earlier, Table 5 shows the results of the findings from each journal database. It can be seen that the relatively small number of results indicates a restriction on open access publications, where most of the publications found have limited access.

TABLE V. RESULT FINDINGS

No	Database Journal	Articles
1	Science Direct	6
2	IEEE Explorer	4
3	MDPI	11
4	Scientific Reports	14

B. *feature fusion methods impact the performance of Vision Transformer Algorithm*

The implementation of feature fusion significantly affects the performance of Vision Transformer (ViT) models in various applications, especially in the fields of image recognition and visual classification, including plant disease detection. Based on the literature review, there are four main effects of feature fusion on ViT model performance. First, feature fusion can significantly improve the accuracy of classification results by combining information from different sources or feature levels. By combining features from different resolutions or domains, the model is able to make more accurate predictions and capture more subtle variations in the data [2], [7], [8], [9], [22], [23]. Secondly, the use of feature fusion can help in improving the feature representation generated by the model, especially on data with high complexity. Feature fusion assists ViT models in identifying and combining disparate information, which may not be visible at first, thus improving generalization and acuity in distinguishing difficult classes [2], [4], [8], [14], [24], [25], [27], [28]. Thirdly, feature fusion methods also play a role in reducing the effects of irrelevant features or noise, thus helping the model to focus on features that are more meaningful and important for the given task [9], [15], [18], [29], [30], [31]. This is especially important in domains such as plant disease detection, where visual features are often highly variable and susceptible to interference from background or poor lighting. Lastly, feature fusion can improve the robustness of ViT models to missing or incomplete data. By utilizing multiple sources of information combined through a feature fusion scheme, the model can still produce reliable predictions even if there are missing or inaccurate pieces of data [7], [31], [32]. These effects show that feature fusion plays an important role in improving the accuracy, robustness, and reliability of Vision Transformer models in various applications, especially in complex and varied tasks.

TABLE VI. THE EFFECTS

No	Effect	References
1	Improve accuracy of classification results by combining information from different sources or feature levels.	[1] [2] [7] [8] [9] [21] [22] [23] [31]
2	The feature can representation generated by the model, especially on data with high complexity	[2] [4] [8] [14] [24] [25] [27] [28] [31]
3	Reducing the effects of irrelevant features or noise	[2] [9] [15] [18] [28] [30] [31]
4	Improve the robustness of ViT models to missing or incomplete data	[7] [31] [32]

C. *The Metrics used to Evaluate The Effectiveness of Vision Transformer*

Evaluating the effectiveness of the feature fusion method with confusion matrix on the Vision Transformer model in classification tasks involves various metrics that assess the accuracy, quality, and robustness of the model. In the confusion matrix method, several important metrics are extracted to understand the performance of the model. Accuracy, for example, is a basic metric that measures the proportion of correct predictions out of the total tested data. Accuracy is often used in previous studies [1] - [8], [14] - [32]. However, accuracy alone is sometimes not enough when the data is not balanced. Therefore, Precision and Recall are also important to consider.

Precision measures the proportion of correct predictions out of all positive predictions [2] - [7], [14], [20]. Which is particularly relevant when false positives are more important to minimize important. In contrast, Recall (True Positive Rate) evaluates how good the model is at detecting true positive elements, especially useful in cases where avoiding false negatives is [2] - [7], [14], [20] - [24]. The combination of Precision and Recall is called F1-Score, which offers a balance between the two and is useful for unbalanced data sets. Kappa, or more commonly known as Cohen's Kappa, measures a statistic used to assess the level of agreement between two independent raters or classifiers. This matrix takes into account the possibility of agreement occurring at random, making it more accurate than a simple measure of agreement [29], [30].

TABLE VII. THE METRICS

No	Metrics	References
1	Accuracy (Acc)	[1] [2] [4] [5] [8] [14] [15] [18] [20] [23] [25] [26] [27] [28] [29] [30] [32]
2	Precision	[2] [5] [7] [14] [20] [22] [23] [24] [28] [32]
3	Recall/ True Positive Rate (TPR)	[2] [5] [7] [14] [23] [22] [24] [26] [28] [30] [32]
4.	Kappa	[29] [30]

D. The Major Challenges in applying Feature Fusion technique with Vision Transformer models

The main challenge in applying feature fusion to the Vision Transformer (ViT) for plant disease detection lies in selecting relevant features from different layers of the model. Not all extracted features make a significant contribution, so proper feature selection is key to improving accuracy and reducing redundancy. Feature selection is complicated by complex interactions between features that can obscure important information [1] - [4], [14] - [32]. One solution to this challenge is the use of a multi-head attention approach that allows the model to account for multiple views of features, resulting in more effective feature incorporation and improved accuracy in plant disease classification [4], [9], [19], [22], [27], [29].

In addition, the challenge of determining the optimal feature weights is also a concern, where isolating the weights for each feature may overlook the contributions from groups of interacting features in the image [5]. To address this, weight regularization approaches and attention mechanisms can be used to prioritize more significant features, reduce the impact of noise, and improve the efficiency of the model in processing complex data [1], [22], [27]. The use of data augmentation techniques also helps improve robustness to variations in environmental conditions, such as lighting and image resolution [7], [22], [30].

Another challenge is the high computational requirements in the application of ViT, especially when handling large and complex datasets [6]. The use of distributed computing frameworks, such as TensorFlow Serving, as well as model pruning methods can help reduce computational requirements without sacrificing model accuracy [19]. With the combination of these solutions, the application of ViT equipped with fusion features becomes more efficient and accurate in detecting plant diseases under diverse field conditions [2], [7], [8], [27], [31].

TABLE VIII. THE MAJOR CHALLENGES

No	Challenges	References
1	Selecting relevant features from different layers of the model, so proper feature selection is key to improving accuracy and reducing redundancy	[1] [2] [4] [5] [8] [14] [15] [18] [22] [23] [25] [27] [28] [29] [30] [32]
2	Using multi-head attention to incorporate multiple views of features	[4] [9] [19] [22] [27] [29]
3	Determining the optimal feature weights and handling feature group interactions	[5]
4	Utilizing weight regularization and attention mechanisms to prioritize significant features	[1] [22] [27]
5	Improving robustness to variations in environmental conditions through data augmentation	[7] [22] [30]
6	Handling high computational requirements when dealing with large and complex datasets	[6]
7	Reducing computational requirements through distributed computing frameworks and model pruning	[19]
8	Achieving more efficient and accurate plant disease detection under diverse field conditions	[2] [7] [8] [27] [31]

IV. CONCLUSION

This study aims to answer three research questions related to the application of feature fusion techniques in Vision Transformer (ViT) models for plant disease detection. First, we found that different feature fusion methods significantly impact the performance of ViT models in terms of accuracy and computational efficiency. These methods help improve classification accuracy by integrating multi-scale features, capturing complex patterns, and reducing noise. Feature fusion techniques also enhance the model's ability to generalize across diverse field conditions, enabling better detection of subtle disease symptoms that may otherwise be missed. Second, in terms of evaluation metrics, previous studies have used several key metrics to assess model performance, including accuracy, precision, recall, F1-score, and Cohen's Kappa. These metrics provide insights into how well the model distinguishes between diseased and healthy plants while ensuring robustness against false positives and negatives.

Third, we identified several major challenges in applying feature fusion to ViT models. These include selecting relevant features from multiple layers, determining optimal feature weights, handling high computational demands, and managing missing or noisy data. Furthermore, there are difficulties in maintaining model accuracy across different environmental conditions, such as varying lighting and image resolution. Solutions such as multi-head attention and weight regularization have been proposed to address these challenges, but further research is needed to refine these approaches. Looking ahead, future research should focus on addressing these challenges to improve the image quality, scalability, efficiency, and accuracy of ViT models in plant disease detection. By overcoming these obstacles, we can enhance the applicability of ViT in real-world agricultural settings, ultimately contributing to more efficient and sustainable farming practices.

ACKNOWLEDGMENT

Thanks are due to Gunadarma University for providing the opportunity and support necessary to carry out this research, as well as to the main promoters and co-promoters for their valuable guidance and assistance.

REFERENCES

[1] Bao, T. Q., Kiet, N. T. T., Dinh, T. Q., & Hiep, H. X. (2020). Plant species identification from leaf patterns using histogram of oriented gradients feature space and convolution neural networks. Ecological Informatics, 67, 101377.

[2] Hossain, S., Tanzim Reza, M., Chakrabarty, A., & Jung, Y. J. (2023). Aggregating different scales of attention on feature variants for tomato leaf disease diagnosis from image data: a transformer driven study. Sensors, 23(7), 3751.

[3] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems.

[4] Wang, J., Yu, X., & Gao, Y. (2021). Feature Fusion Vision Transformer for Fine-Grained Visual Categorization. International Journal of Computer Vision.

[5] Lye, H. Z. M., & Ng, K. W. (2023). Processing Plant Diseases Using Transformer Model. JOIV: International Journal on Informatics Visualization, 7(4), 2550-2557.

[6] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., & Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. International Conference on Learning Representations.

[7] Zhu, D., Tan, J., Wu, C., Yung, K., & Ip, A. W. (2023). Crop disease identification by fusing multiscale convolution and vision transformer. Sensors, 23(13), 6015.

[8] Yong, W. C., Ng, K. W., Haw, S. C., Naveen, P., & Ng, S. B. (2024). Leaf Condition Analysis Using Convolutional Neural Network and Vision Transformer. International Journal of Computing and Digital Systems, 16(1), 1-10.

[9] Chen, Y., Wang, X., Chen, Z., Wang, K., Sun, Y., Jiang, J., & Liu, X. (2023). Classification of Camellia oleifera Diseases in Complex Environments by Attention and Multi-Dimensional Feature Fusion Neural Network. Plant Science.

[10] A. Carrera-Rivera, W. Ochoa, F. Larrinaga, and G. Lasa, "How-to conduct a systematic literature review: A quick guide for computer science research," MethodsX, pp. 1–12, Nov. 2022, doi: 10.1016/j.compind.2022.103730.

[11] C. Kurniawan, Y. Rosmansyah, and B. Dabarsyah, "A Systematic Literature Review on Virtual Reality for Learning," in 2019 IEEE 5th International Conference on Wireless and Telematics (ICWT), IEEE, 2019.

[12] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén, "Systematic Literature Reviews," in Experimentation in Software Engineering, Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 45–54. doi: 10.1007/978-3-642-29044-2_4.

[13] J. Bettany-Saltikove, How to do a systematic literature review in nursing: a step-by-step guide. Open University Press, 2012.

[14] Hosny, K. M., El-Hady, W. M., Samy, F. M., Vrochidou, E., & Papakostas, G. A. (2023). Multi-class classification of plant leaf diseases using feature fusion of deep convolutional neural network and local binary pattern. IEEE Access, 11, 62307-62317.

[15] Önler, E. (2023). Feature fusion based artificial neural network model for disease detection of bean leaves. Electronic Research Archive.

[16] Jiang, S., Xu, Y., Li, D., & Fan, R. (2022). Multi-scale fusion for RGB-D indoor semantic segmentation. Scientific Reports, 12(1), 20305.

[17] Wang, X., Wu, Y., Ming, Y., & Lv, H. (2020). Remote sensing imagery super resolution based on adaptive multi-scale feature fusion network. Sensors, 20(4), 1142.

[18] Wang, S., Qu, J., Zhang, Y., & Zhang, Y. (2023). Multimodal emotion recognition from EEG signals and facial expressions. IEEE Access, 11, 33061-33068.

[19] Gao, R., Dong, Z., Wang, Y., Cui, Z., Ye, M., Dong, B., ... & Yan, S. (2024). Intelligent cotton Pest and disease detection: Edge computing solutions with transformer technology and knowledge graphs. Agriculture, 14(2), 247.

[20] Soh, K. S., Moung, E. G., Danker, K. J. J., Dargham, J. A., & Farzamnia, A. (2024). Cocoa Diseases Classification using Deep Learning Algorithm. In ITM Web of Conferences (Vol. 63, p. 01014). EDP Sciences.

[21] Borhani, Y., Khoramdel, J., & Najafi, E. (2022). A deep learning based approach for automated plant disease classification using vision transformer. Scientific Reports, 12(1), 11554.

[22] Chen, Z., Zhou, H., Lin, H., & Bai, D. (2024). TeaViTNet: Tea Disease and Pest Detection Model Based on Fused Multiscale Attention. Agronomy, 14(3), 633.

[23] Faisal, M., Leu, J. S., & Darmawan, J. T. (2023). Model Selection of Hybrid Feature Fusion for Coffee Leaf Disease Classification. IEEE Access, 11, 62281–62291. https://doi.org/10.1109/ACCESS.2023.3286935

[24] Li, G., Wang, Y., Zhao, Q., Yuan, P., & Chang, B. (2023). PMVT: a lightweight vision transformer for plant disease identification on mobile devices. Frontiers in Plant Science, 14, 1256773.

[25] Li, X., & Li, S. (2022). Transformer help CNN see better: a lightweight hybrid apple disease identification model based on transformers. Agriculture, 12(6), 884.

[26] Barman, U., Sarma, P., Rahman, M., Deka, V., Lahkar, S., Sharma, V., & Saikia, M. J. (2024). Vit-SmartAgri: vision transformer and smartphone-based plant disease detection for smart agriculture. Agronomy, 14(2), 327.

[27] Tabbakh, A., & Barpanda, S. S. (2023). A Deep Features extraction model based on the Transfer learning model and vision transformer" TLMViT" for Plant Disease Classification. IEEE Access.

[28] Yang, B., Li, M., Li, F., Wang, Y., Liang, Q., Zhao, R., ... & Wang, J. (2024). A novel plant type, leaf disease and severity identification framework using CNN and transformer with multi-label method. Scientific Reports, 14(1), 11664.

[29] Lu, X., Yang, R., Zhou, J., Jiao, J., Liu, F., Liu, Y., ... & Gu, P. (2022). A hybrid model of ghost-convolution enlightened transformer for effective diagnosis of grape leaf disease and pest. Journal of King Saud University-Computer and Information Sciences, 34(5), 1755-1767.

[30] Sun, C., Zhou, X., Zhang, M., & Qin, A. (2023). SE-VisionTransformer: Hybrid Network for Diagnosing Sugarcane Leaf Diseases Based on Attention Mechanism. Sensors, 23(20), 8529.

[31] Chen, X., Pan, J., Lu, J., Fan, Z., & Li, H. (2023, June). Hybrid cnn-transformer feature fusion for single image deraining. In Proceedings of the AAAI conference on artificial intelligence (Vol. 37, No. 1, pp. 378-386).

[32] Chen, Z., Wang, G., Lv, T., & Zhang, X. (2024). Using a Hybrid Convolutional Neural Network with a Transformer Model for Tomato Leaf Disease Detection. Agronomy, 14(4), 673.