CHAPTER 2

LITERATURE SURVEY

2.1 Introduction to ViT

Vision Transformers (ViTs) represent a major step forward in computer vision. They change how machines understand and interpret images. Building on the success of transformers in natural language processing, ViTs use the same attention-based method for visual data. Instead of analyzing an image with traditional convolutional layers, they break it into small patches, turn these patches into sequences, and input them into a transformer that learns relationships between different visual areas. The main strength of ViTs is their self-attention mechanism. This feature lets the model capture long-range dependencies across the whole image. Unlike convolutional neural networks (CNNs), which handle information locally within a fixed receptive field, ViTs can focus on multiple regions at once. This ability helps them balance fine details with the overall structure of an image, leading to excellent results in tasks like image classification, segmentation, and object detection.

Typically, Vision Transformers (ViTs) begin with pre-training on large image datasets like ImageNet or JFT. During this phase, they learn various visual patterns and representations. This pre-training gives them a strong base for understanding general visual ideas. Later, they can be fine-tuned for specific tasks. Because of this broad learning process, ViTs often achieve higher accuracy and flexibility than traditional CNNs in many computer vision applications. They can be fine-tuned for specific tasks later on. Because of this wide-ranging learning process, ViTs often achieve better accuracy and adaptability than traditional CNNs in many computer vision applications. However, they require significant computational

power and memory. This makes it hard to use them on low-power or mobile devices. To solve these problems, researchers have developed more efficient hybrid models, including MobileViT, EfficientFormer, and MobilePlantViT. These models mix convolutional layers with transformer components. They maintain the efficiency of CNNs while also using the deep understanding provided by transformers. As a result, Vision Transformers are increasingly used not just in traditional computer vision tasks but also in specialized areas like plant disease detection, medical imaging, and remote sensing, where capturing subtle visual details is crucial.

The future for ViTs focuses improving their efficiency, clarity, and strength in facing real-world challenges such as noise and changing lighting. As Vision Transformers keep evolving, they will become some of the most powerful tools in visual recognition. They help bridge the gap between human perception and machine understanding.

2.1 Existing Works

Kai Han, Yunhe Wang, Qi Tian, [1] and their team introduce GhostNet, a simple convolutional neural network architecture that creates more feature maps while reducing computational costs. The main idea behind GhostNet is the Ghost module. It uses standard convolution to create essential feature maps. Then, it creates extra “ghost” feature maps using simple linear changes. This method significantly lowers computational costs and the number of model parameters while keeping accuracy similar to that of more complex models. GhostNet reaches a top-1 accuracy of 75.7% on ImageNet with just 141 million FLOPs. It outperforms other lightweight architectures like MobileNetV3 in terms of efficiency. The results show that CNNs can effectively use feature redundancy to build compact and fast architectures. This makes GhostNet a strong foundation for mobile and embedded vision solutions.

Qibin Hou, Daquan Zhou, and Jiashi Feng [2] present Coordinate Attention, a useful attention mechanism that adds positional information to channel attention in convolutional neural networks. Unlike earlier modules like CBAM or SE blocks, Coord Attention splits global pooling into two one-dimensional feature encodings across height and width. This method captures long-range dependencies while maintaining location sensitivity. It allows the network to focus on both the "what" and "where" of features, improving feature refinement. Experiments on ImageNet and various object detection benchmarks demonstrate that Coord Attention raises top-1 accuracy by 0.8% with minimal additional computation. The authors indicate that this technique benefits lightweight models such as MobileNetV2 and ShuffleNetV2, highlighting its value in mobile and real-time computer vision applications.

Li, Song, Zhang, Wang, and their research team [3] introduce EfficientFormer. This is a vision transformer framework that combines the accuracy of transformer models with the speed and efficiency of MobileNet. The authors create a hybrid design that includes Fused-Inverted Residual (Fused-IR) blocks and linear attention. This method lowers computational costs while still enabling global feature extraction. By modifying layers and reducing tokens at different stages, EfficientFormer achieves a good balance between accuracy and efficiency. It reaches up to 79.2% Top-1 accuracy on ImageNet with just 1.6 GFLOPs. This research demonstrates that using methods for convolutional efficiency in transformer designs can improve performance on edge and mobile devices. The study merges the compactness of CNNs with the strong representation of ViTs, prompting further exploration of lightweight transformer architectures for practical use.

Xu, Wang, and Han et al. [4] introduce Linear Differential Attention (LiDA), a new attention method that reevaluates traditional linear attention by using differential

operators. The authors address the weaknesses of current linear attention methods, which often fail to match the expressiveness of full attention. LiDA enhances representational ability by using differential kernels that effectively capture complex feature interactions while maintaining linear computational efficiency. Experimental results on the ImageNet-1K and CIFAR benchmarks show that LiDA outperforms earlier linear transformers like Performer and Linear Transformer by up to 1.8% in Top-1 accuracy, all while using fewer parameters. This research highlights the potential of differential-based feature mixing for building strong, noise-resistant, and efficient vision transformers. It represents a major advancement in creating scalable ViT architectures that are suitable for devices with limited resources.

Zhuang Liu, Hugo Touvron, and their team [5] present ConvNeXt V2, a new convolutional architecture that connects classic ConvNets with modern vision transformers. They build on the original ConvNeXt by adding Bottleneck Feed-Forward Networks (FFNs) and using masked autoencoder (MAE) pretraining to improve efficiency and learning. ConvNeXt V2 achieves impressive accuracy while significantly lowering computational costs by using inverted bottleneck designs and better normalization methods. In tests on ImageNet-1K, ConvNeXt V2 reaches a Top-1 accuracy of 86.8%, outperforming earlier ConvNet-based models. The paper shows that with thoughtful architectural choices and combined training, convolutional models can match or even surpass the accuracy and scalability of transformer models. This research offers a practical framework for developing compact, high-performing networks that are suitable for various vision tasks.

Sachin Mehta and Mohammad Rastegari [6] introduce MobileViT, a straightforward hybrid design that combines the strengths of convolutional neural networks (CNNs) and vision transformers (ViTs). This design targets mobile and edge devices. The authors propose using transformer blocks instead of large convolutional layers. These blocks capture long-range relationships while keeping

local spatial details through convolutional methods. MobileViT divides images into non-overlapping patches and uses multi-head self-attention for processing. It then reintegrates these patches into the convolutional framework. This allows it to recognize both global and local features. In tests on the ImageNet dataset, MobileViT-XXS achieves a Top-1 accuracy of 74.8% with only 0.95M parameters. This model outperforms networks like MobileNetV2 and EfficientNet-B0, all while keeping similar or lower computational costs. The paper demonstrates that combining transformers with lightweight CNNs significantly boosts representational power without losing efficiency. This research opens the door for hybrid models like MobileViTv2 and MobilePlantViT, highlighting an important step in creating efficient vision models.

Sachin Mehta and Mohammad Rastegari [7], in their publication titled "MobileViTv2: Separable Self-Attention for Mobile Vision Transformers" (Trans. Mach. Learn. Res., 2023), introduce an improved hybrid model aimed at enhancing the efficacy of transformers on mobile and embedded platforms. Building upon the concepts of MobileViTv1, the authors present Separable Linear Self-Attention, a technique that offers greater computational efficiency than traditional self-attention by splitting the projections of keys and values into distinct linear transformations. This adjustment greatly diminishes both the number of parameters and latency while preserving high accuracy in classification and detection tasks. Additionally, the model incorporates lightweight convolutional stems to enhance token mixing and more effectively capture spatial features. Experimental findings on ImageNet-1K indicate that MobileViTv2-S achieves a Top-1 accuracy of 78.1% with nearly 25% fewer FLOPs than MobileViTv1, thereby establishing a new equilibrium between efficiency and accuracy. The paper underscores that separable self-attention enables the efficient scaling of transformers on hardware with limited resources. This advancement marks a significant step forward in developing ViT architectures that are suitable for mobile use while maintaining the benefits of a global receptive field typically

found in full-sized transformers.

Wasi Ullah, M. Irfan, M. A. Khan, and A. Hussain [8], introduce a lightweight vision transformer for detecting agricultural diseases in their study titled “AppViT: Efficient Identification and Classification of Apple Leaf Diseases using Lightweight Vision Transformer” (Discover Sustainability, Springer Nature, 2024). The authors combine convolutional layers that capture local features with transformer encoders that handle global relationships. This creates a hybrid model for analyzing plant leaf images. The architecture employs depthwise separable convolutions and attention layers with reduced dimensions, achieving high accuracy while maintaining low computational needs. When tested on various datasets related to apple leaf diseases, AppViT demonstrates over 98% accuracy in classification and maintains a minimal parameter count. It outperforms CNN-based models like EfficientNet and ResNet in both accuracy and inference speed. The results indicate that transformer-based models can effectively address complex problems in plant pathology, even under different environmental conditions. This research highlights the growing importance of lightweight transformer models in precision agriculture and paves the way for potential applications of ViTs in crop monitoring and disease management.

Thakur, A., and Khanna, A. [9] introduce a lightweight vision transformer for detecting agricultural diseases in their study titled “AppViT: Efficient Identification and Classification of Apple Leaf Diseases using Lightweight Vision Transformer” (Discover Sustainability, Springer Nature, 2024). The authors combine convolutional layers that capture local features with transformer encoders that handle global relationships. This creates a hybrid model for analyzing plant leaf images. The architecture employs depthwise separable convolutions and attention layers with reduced dimensions, achieving high accuracy while maintaining low computational needs. When tested on various datasets related to apple leaf diseases, AppViT demonstrates over 98% accuracy in classification and maintains

a minimal parameter count. It outperforms CNN-based models like EfficientNet and ResNet in both accuracy and inference speed. The results indicate that transformer-based models can effectively address complex problems in plant pathology, even under different environmental conditions. This research highlights the growing importance of lightweight transformer models in precision agriculture and paves the way for potential applications of ViTs in crop monitoring and disease management.

Kumar, S., and Arya, D. [10], in their work titled “PDLCViT: Plant Disease Localization and Classification via Vision Transformer” (Springer, 2024), introduce a comprehensive framework aimed at detecting and pinpointing diseases on plant leaves through the use of Vision Transformers (ViTs). The authors developed PDLCViT to harness the self-attention mechanism, allowing it to simultaneously achieve accurate classification and localization of affected areas without the need for specific bounding-box annotations. This model combines a patch-based ViT encoder with a simplified decoder that creates class activation maps, which effectively identify diseased regions on plant leaves. Experimental findings from various datasets of plant diseases indicate that PDLCViT attains a classification accuracy of 98.3% and a localization precision of 96.7%, surpassing CNN-based detection models like ResNet50 and InceptionV3. Furthermore, attention visualization shows that the model successfully concentrates on relevant lesion areas, even when faced with occlusion or noise. The authors emphasize the importance of transformer-based localization in agriculture, highlighting that PDLCViT provides both diagnostic insights and precise visual reasoning. This research represents a significant advancement in the explainable and automated assessment of plant diseases through ViT-based methodologies.

2.2 References

[1] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, “GhostNet: More Features from Cheap Operations,” in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition
(CVPR)*, 2020, pp. 1580–1589. https://doi.org/10.48550/arXiv.1911.11907

[2] Q. Hou, D. Zhou, and J. Feng, “Coordinate Attention for Efficient Mobile Network Design,” in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 13713–13722.
https://doi.org/10.48550/arXiv.2103.02907

[3] L. Li, X. Song, Z. Zhang, Y. Li, L. Wang, and H. Hu, “EfficientFormer: Vision Transformers at MobileNet Speed,” in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2022. https://doi.org/10.48550/arXiv.2206.01191

[4] X. Xu, Y. Wang, and K. Han, “Linear Differential Attention: Rethinking Linear Attention for Vision Transformers,” in *Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, 2024.DOI: 10.1007/978-3-031-50215-6_13,
https://link.springer.com/chapter/10.1007/978-3-031-50215-6_13

[5] Z. Liu, H. Touvron, et al., “ConvNeXt V2: Co-Designing and Scaling ConvNets with Masked Autoencoders,” in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023. https://doi.org/10.48550/arXiv.2301.00808

[6] S. Mehta and M. Rastegari, “MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer,” in *Proc. Int. Conf. on Learning Representations (ICLR)*, 2022. https://doi.org/10.48550/arXiv.2110.02178

[7] S. Mehta and M. Rastegari, “MobileViTv2: Separable Self-Attention for Mobile Vision Transformers,” *Trans. Machine Learning Research (TMLR)*, 2023.
https://openreview.net/forum?id=tBl4yBEjKi.

[8] W. Ullah, M. Irfan, M. A. Khan, and A. Hussain, “AppViT: Efficient Identification and Classification of Apple Leaf Diseases using Lightweight Vision Transformer,” *Discover Sustainability*, Springer, vol. 4, no. 3, pp. 1–15, 2024. https://link.springer.com/article/10.1007/s43621-024-00307-1

[9] A. Thakur and A. Khanna, “PlantXViT: Explainable Vision Transformer Enabled CNN for Plant Disease Identification,” *arXiv preprint arXiv:2207.07919*, 2022. https://doi.org/10.48550/arXiv.2207.07919

[10] S. Kumar and D. Arya, “PDLCViT: Plant Disease Localization and Classification via Vision Transformer,” *Springer Journal of Signal, Image and Video Processing*, w2024. https://link.springer.com/article/10.1007/s44196-024-00597-3