Introduction

As a transformative architecture in computer vision, the Swin Transformer has shown itself to be exceptionally good at effectively capturing both local and global dependencies. Although they have been the industry standard for a long time, traditional convolutional neural networks (CNNs) have limited capacity to represent global information because of their small receptive fields [1]. Furthermore, CNNs frequently need a lot of processing power, especially when used for complicated image identification tasks [2]. By applying attention within non-overlapping local windows and connecting these windows to record global interactions, the Swin Transformer proposes a hierarchical vision transformer design that addresses these issues. Tasks including object detection, instance segmentation, and image classification are made much more efficient and perform better using this method [3].

A noteworthy use of the Swin Transformer is in the Swin-MLP technique, which combines the Swin Transformer with a multi-layer perceptron (MLP) to identify the look quality of strawberries. Enhancing market value of strawberries requires accurate strawberry quality identification, however manual sorting methods are frequently labor-intensive and prone to human mistake [1]. This is addressed by Swin-MLP, which uses an MLP for classification after extracting picture features using the Swin Transformer. This approach outperforms conventional CNN-based models in terms of training speed and accuracy, achieving a classification accuracy of 98.45%. This shows the versatility of the Swin Transformer in applications connected to agriculture [1].

Swin-Depth, a depth estimate model that makes use of the hierarchical representation learning of Swin Transformer, is another noteworthy advancement. In domains like autonomous driving and robot navigation, the

ability to estimate depth through monocular sensors is essential. Nevertheless, computational complexity and restricted global context modeling are common problems for conventional CNN-based techniques [2]. By including a multi-scale fusion attention mechanism, Swin-Depth addresses these problems by effectively capturing both local and global depth information while requiring fewer parameters. On datasets such as KITTI and NYU, this model has outperformed previous CNN-based depth estimation methods, achieving state-of-the-art performance [3].

Swin Transformers are unique in that they may be used for a variety of applications, such as depth estimation and agricultural quality monitoring, with no computing overhead. They are an effective tool for contemporary computer vision problems due to their capacity to represent intricate spatial relationships and preserve high-resolution detail [2]. Swin Transformer is a significant advancement in the accuracy and efficiency of deep learning models, as shown in several domains such as robotics and image recognition [1], [3].

- **[1]** refers to the paper titled *"Swin-MLP: A Strawberry Appearance Quality Identification Method by Swin Transformer and Multi-Layer Perceptron"*.
- **[2]** refers to the paper titled *"Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows"*.
- **[3]** refers to the paper titled *"Swin-Depth: Using Transformers and Multi-Scale Fusion for Monocular-Based Depth Estimation"*.