Hamba: Single-view 3D Hand Reconstruction with Graph-guided Bi-Scanning Mamba

Paper

The Hamba model estimates 3D hand reconstruction from a single RGB image, through a new approach of combining graph learning and state-space modeling. It introduces a Graph-guided State Space block to reduce token usage compared to transformer/attention-based methods.

Motivation

Existing state-of-the-art transformer-based methods require excessive computational resources, due to attention-based methods using quadratic speed and memory. Hamba resolves this issue by integrating graph learning with the state-space modeling to optimize token usage and bring Mamba's speed and memory advantages to 3D hand reconstruction.

Architecture

Backbone

The image is fed into a Vision Transformer backbone to extract a set amount of tokens, which are then downsampled using convolution layers.

Token Sampler

The 2D hand joints are initially predicted by the Joint Regressor, which consists of stacked 2D Selective Scan (SS2D) blocks and an MLP head to regress the initial MANO parameters. The Joint Regressor first regresses the 3D joints and then projects them back to the 2D image plane with a predicted camera translation and a predefined focal length. The tokens are then aligned with the 2D joints with bilinear interpolation.

Graph-guided State Space (GSS) Block

Bidirectional scans are performed over the sampled joint tokens to model both local and global join dependencies for hand mesh reconstruction, using a Semantic Graph Convolution Network (GCN) block. The relationships between hand joints are modeled with a GCN using a predefined graph structure based on the hand joint skeleton.

Fusion Module

The features from the GSS block, global token mean, and joints, are combined and the MANO parameters are regressed them to generate the 3D hand keypoints and mesh.

Limitations

1. The model does not have temporal feature extraction and processes frames in a video independently.
2. While the majority of the model does not rely on attention mechanisms or transformers, the backbone still uses a Vision Transformer, which is computationally intensive.