MambaPose:
Low Memory Footprint 3D Human Pose Estimation

Elif Gurkan (1), Melih Darcan (2), Nazli Ikizler Cinbis (1,2)

1 Hacettepe University Computer Science Department
2 Hacettepe University Artificial Intelligence Department

Introduction

MambaPose represents a breakthrough in 3D human pose estimation, addressing critical challenges in computer vision with a low memory footprint for practical application in real-time environments.

Traditional pose estimation methods struggle with real-time processing, high computational demands, and inaccuracies in complex scenes.

Our project introduces a novel, integrated approach leveraging state-of-the-art technologies:

- RAFT for optical flow estimation,
- Depth Anything for depth estimation,
- and MambaPose for advanced pose estimation.

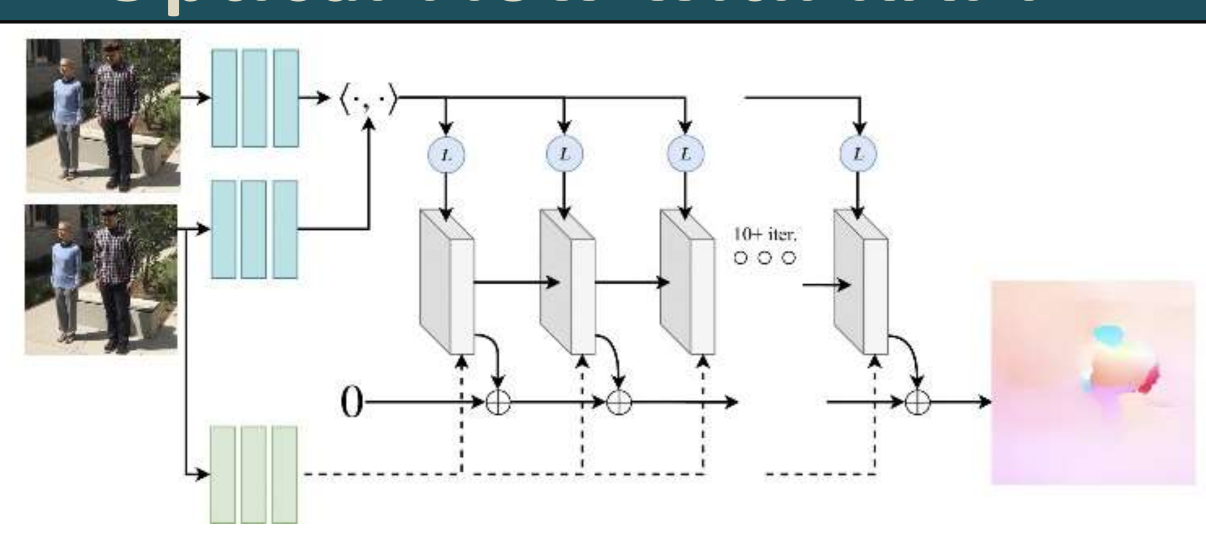
This synergistic integration enhances the accuracy and efficiency of detecting and analyzing human poses, critical for applications ranging from augmented reality to surveillance.

By optimizing these techniques, MambaPose offers a promising solution for dynamic and accurate 3D human pose estimation in various real-world scenarios.

Depth Anything

RAFT: Recurrent All-Pairs Field Transforms for Optical Flow

Optical Flow with RAFT



RAFT is a deep learning model used for <u>high-accuracy and efficient optical flow estimation</u>. This method predicts the movement of pixels between two consecutive images by considering all pairs of pixels and modeling the relationships between these pairs through a series of recurrent updates.

The model consists of <u>three main parts</u>:

Feature Extraction: Two consecutive images are transformed into features using a pre-trained CNN.

Correlation: The extracted features are combined to create a correlation volume for all pairs of pixels.

Flow Estimation: An initially random flow estimate is iteratively refined using the correlation volume and features. Each iteration further refines the estimate.

RAFT is particularly <u>significant in the field of 3D pose estimation for several reasons</u>:

Accuracy in Motion Detection: RAFT's pixel-wise motion predictions are crucial for reconstructing precise 3D poses from sequences of images, enhancing the accuracy of the pose estimation.

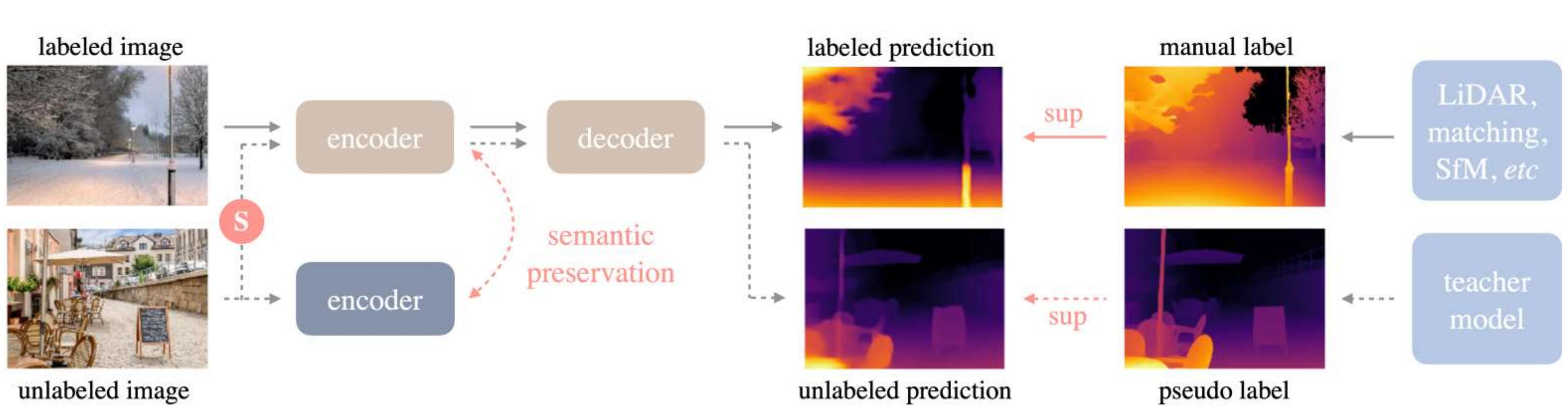
Handling Complex Motions: The iterative refinement process allows RAFT to excel in scenes with complex motions and varying dynamics, which are common in applications where 3D pose estimation is crucial.

Temporal Consistency: The recurrent nature of RAFT helps maintain consistency in optical flow across frames, ensuring the continuity and reliability of 3D pose trajectories.

Robustness to Varied Conditions: RAFT's effectiveness in diverse filming conditions ensures that 3D pose estimation remains reliable under various scenarios, such as different lighting and weather conditions or rapid movements.

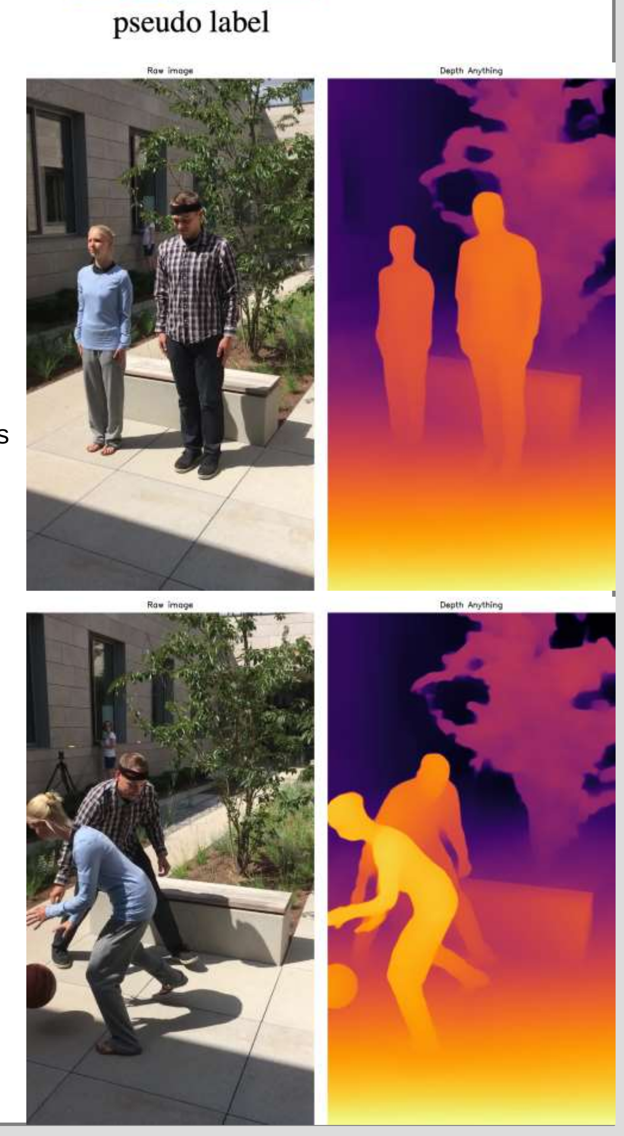
RAFT performs exceptionally well in challenging and complex scenes, delivering consistent results even in varying conditions. It is suited for video processing, motion analysis, and real-time applications, making it invaluable in fields such as augmented reality, virtual reality, advanced robotics, and clinical gait analysis.

Depth Estimation with Depth-Anything

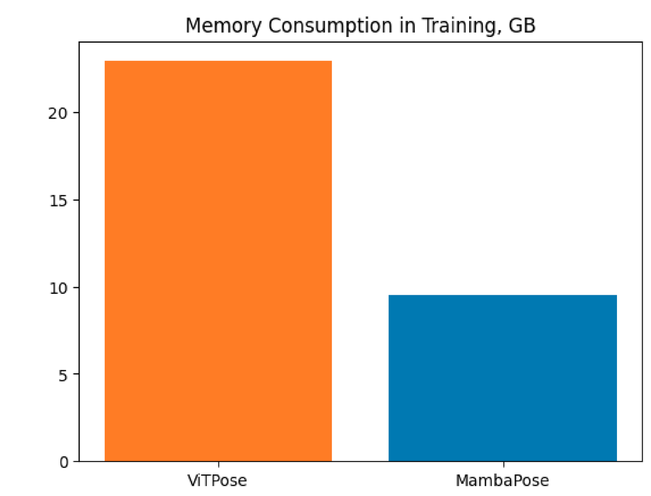


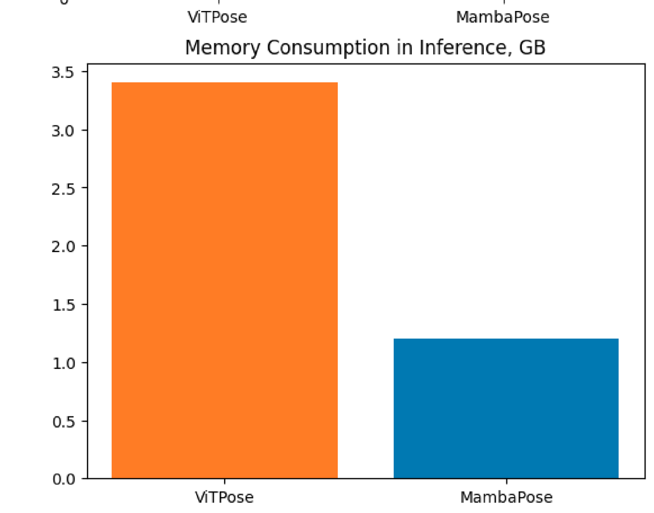
The "Depth Anything" model <u>is critically important for 3D pose estimation</u> due to its ability to generate accurate depth maps from single images. Depth maps are fundamental in converting 2D image data into 3D spatial information, which is crucial for understanding the pose and orientation of objects and humans in three-dimensional space. Here's how the "Depth Anything" model enhances 3D pose estimation:

- **Depth Precision**: The model's capability to provide precise depth estimates helps in accurately determining the position of objects or body parts in space, which is essential for calculating their 3D poses from 2D images.
- **Contextual Understanding**: By integrating depth estimation with semantic information, the "Depth Anything" model aids in distinguishing between foreground and background elements. This separation is crucial for identifying relevant objects or figures for pose estimation, especially in cluttered or complex scenes.
- **Enhanced Accuracy in Complex Environments**: The robustness of the model in diverse and challenging conditions, such as poor lighting or occluded environments, ensures that the depth estimation and consequently the 3D pose estimation remain reliable. This is essential for applications that operate in real-world conditions where such challenges are common.
- **Integration with Other Modalities**: The ability to integrate the depth estimation from "Depth Anything" with other modalities, like optical flow or additional sensor data (e.g., LiDAR), further enriches the 3D pose estimation process. This multimodal approach can lead to more accurate and robust pose estimations, mitigating the limitations of using single image data alone.
- **Temporal Consistency**: In dynamic scenarios where continuous pose estimation is required, the model's ability to maintain depth consistency across frames supports stable and continuous tracking of 3D poses. This consistency is vital for applications such as motion capture and gesture recognition, where tracking the movement over time is necessary.



MambaPose





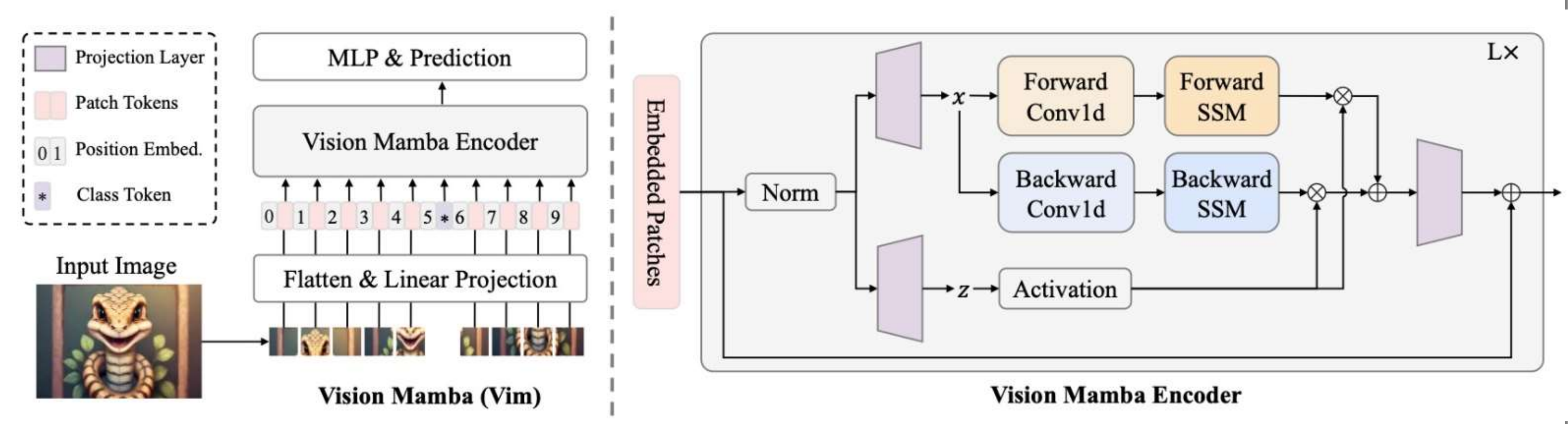
MambaPose innovatively <u>combines RAFT for detailed motion capture with Depth Anything for accurate depth perception</u>, setting a new standard in 3D human pose estimation. This hybrid model architecture is optimized for high performance in real-time applications, even on devices with limited computational resources. Key Features:

Efficient Memory Usage:

- Substantially lower memory requirements during both training and inference phases compared to ViTPose, facilitating deployment in memory-constrained environments.
- Memory Consumption Graphs illustrate the stark contrast in resource utilization, emphasizing MambaPose's suitability for real-time applications.

High Accuracy and Real-Time Processing:

- Incorporates state-of-the-art optical flow and depth estimation technologies to achieve high precision in dynamic and complex scenes.
- Designed for real-time processing, crucial for applications in augmented reality, interactive gaming, and clinical analysis.



Innovative Architecture of MambaPose

MambaPose employs <u>a bidirectional approach with State Space Models (SSMs)</u> to enhance the accuracy and efficiency of 3D human pose estimation. This architecture allows the model to leverage temporal context from both past and future frames, providing a comprehensive understanding of dynamic scenes.

<u>Key Features of Bidirectional SSMs</u>:

- Dual Direction Processing: Unlike traditional models that process data in a single direction, MambaPose processes information both forward and backward. This ensures a richer context, improving the accuracy of pose predictions by incorporating insights from the sequence's history and future projections.
- State Space Framework: Adapts from control theory, the SSM framework in MambaPose describes the pose states and their transitions based on observed movements. This method effectively captures the dynamics of human motion, facilitating precise pose estimation across varied activities.
- Enhanced Generalization: By processing sequences in both directions, MambaPose can generalize better across different environments and activities. This is particularly useful in complex scenarios where understanding the entire sequence context is crucial for accurate predictions.

<u>Efficiency and Robustness</u>:

- The architecture's efficiency is depicted in the Vision Mamba Encoder diagram, showing streamlined data processing through multiple encoder layers. This design minimizes computational load while maximizing the extraction of relevant features, making MambaPose ideal for real-time applications.

Results & Impact of MambaPose

Training and Validation Dynamics

Rapid Model Convergence:

MambaPose's training dynamics for both 2D and 3D models show rapid convergence, indicating the model's robust learning algorithms that effectively capture the complexities of human motion without overfitting. The training loss decreases sharply within initial epochs and then stabilizes, which highlights efficient learning mechanisms.

Stable Validation Loss:

The stability of the validation loss across training epochs underscores the model's ability to generalize well on unseen data. This consistency is essential for deploying the model in real-world settings, where varied and unpredictable human movements are common.

Enhanced Accuracy with Lower Resource Usage

Memory Consumption:

Comparative graphs illustrate that MambaPose consumes significantly less memory than traditional models like ViTPose during both training and inference phases. This optimized memory usage ensures that MambaPose can be deployed on edge devices where resource constraints are a critical concern.

Efficient Processing:

The architectural efficiencies built into MambaPose allow it to process data faster and with lower computational overhead. This is particularly advantageous in scenarios requiring real-time data processing.