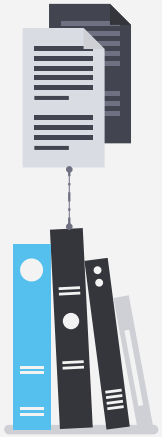
Combine MHFormer and Token Pruning Cluster for observing efficiency and effectiveness

DIP Final Project Team 15

鍾名捷 313551130
洪子奇 313551159
葉軒宇 313553036
李東諺 313553052

Table of contents



Introduction

MHFormer : Multi-Hypothesis Transformer for 3D Human Pose Estimation, The method used is to synthesize the final 3D pose by using multiple hypothetical poses, It is intended to solve the problem of estimating 3D human poses from monocular videos faces depth ambiguity and self-occlusion.

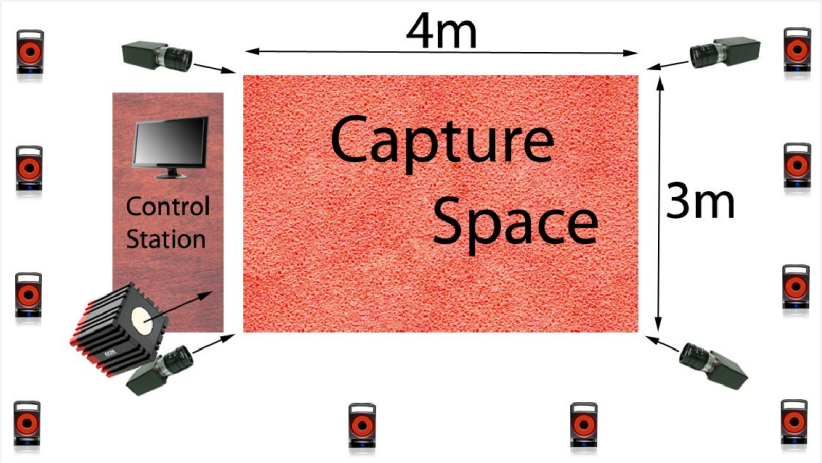
TPC (Token Pruning Cluster) : It is used to solve the problem of high computing requirements of the Video Pose Transformer (VPT). Reduce the amount of calculation by pruning less important tokens.

Related work

- **Single-View 3D Human Pose Estimation**
  - Traditionally divided into two subtasks: 2D pose detection and 2D-to-3D lifting.
- **Vision Transformers**
  - The adoption of Transformers, such as ViT for image classification and PoseFormer for human pose estimation, has shown significant improvements due to their ability to capture long-range dependencies.
- **Multi-Hypothesis Methods**
  - Recent works have started generating multiple hypotheses to tackle the inverse nature of the 3D pose estimation problem.

Dataset/Platform

Human3.6M : The Human3.6M dataset contains 3.6 million images from 11 subjects performing the following 15 behaviors (directions, discussion,sitting on chair, smoking,making purchases, greeting, waiting, taking photo, talk by phone, walking, walking dog, walking together, posing, eating, activities while seated) , which were recorded from multiple perspectives. These images are annotated with the positions of 3D human joints, that is, there are 3D coordinates on each image.



Baseline

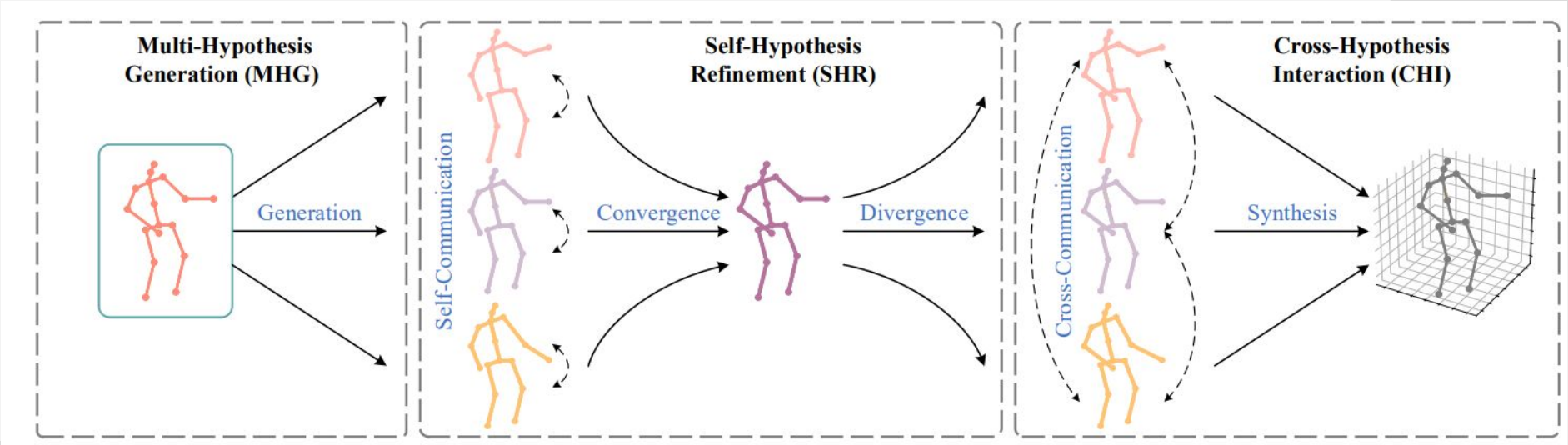
MHFormer is built on three main modules, forming a three-stage processing pipeline:

1. **Multi-Hypothesis Generation (MHG)**
2. **Self-Hypothesis Refinement (SHR)**
3. **Cross-Hypothesis Interaction (CHI)**

In addition to these three main modules, MHFormer includes two auxiliary modules:

1. **Temporal Embedding**
2. **Regression Head**

Baseline


Multi-Hypothesis
Generation (MHG)
Generation
Self-Hypothesis
Refinement (SHR)
Self-Communication
Convergence
Divergence
Cross-Hypothesis
Interaction (CHI)
Cross-Communication
Synthesis

Main approach

Main idea of Token Pruning Cluster

Token Pruning Cluster (TPC) is one of the core modules proposed in our project. It focuses on dynamically selecting the most representative pose tokens in the deeper structures of a Transformer.

Problem Background

In Transformers, each video frame is treated as a token, and processing long sequences (e.g., 243 frames) incurs high computational costs.

TPC's Goal

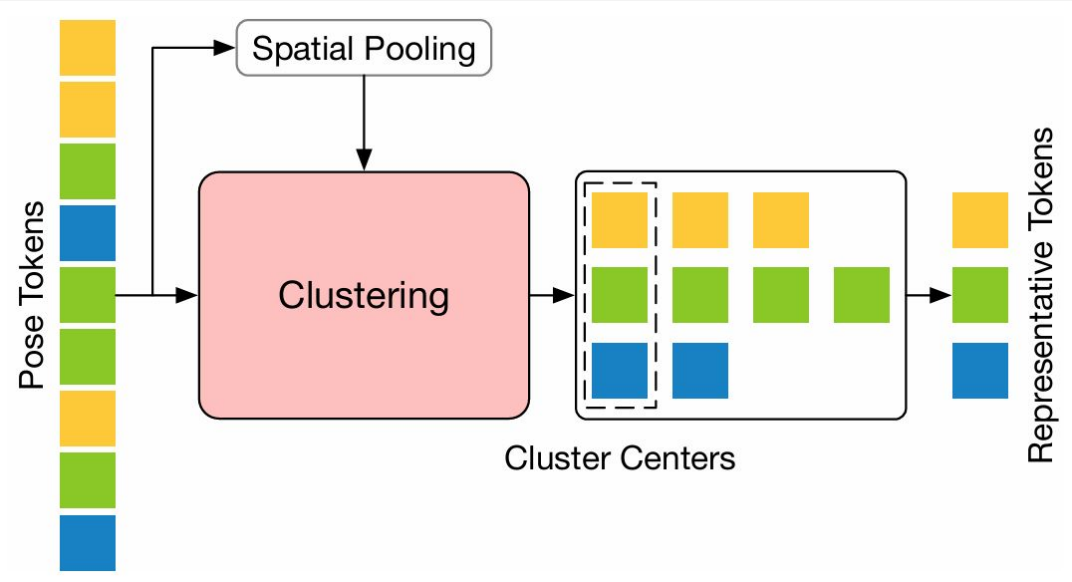
Dynamically prune redundant frames while retaining a few representative ones to reduce computational costs without compromising model accuracy.

Main approach

The design of TPC including 3 steps:

1. **Spatial Pooling**
   Remove spatial redundancy and focus on selecting features along the temporal dimension.
2. **Density Peaks Clustering**
   Cluster frames based on feature similarity and select cluster centers as representative frames.
3. **Selecting Representative Frames**
   After clustering, the representative frames are selected for downstream processing.

Main approach



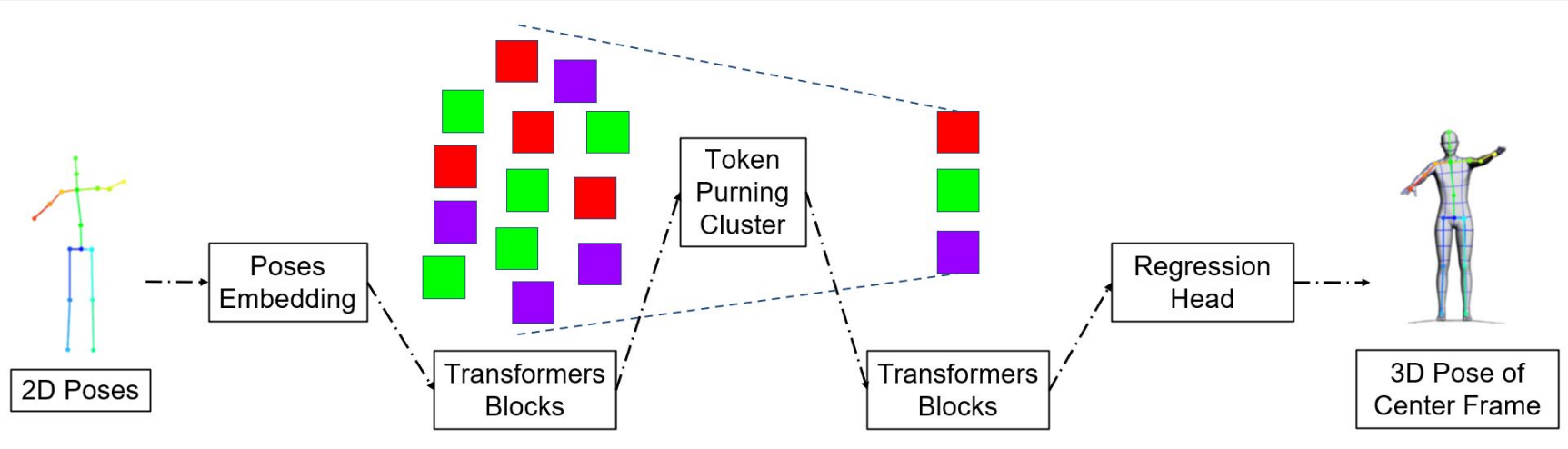
Main approach

Advantages of TPC

1. **Dynamism:**
   Unlike static uniform sampling methods, TPC dynamically selects representative frames based on feature distribution, making it better suited to the data's characteristics
2. **Semantic Diversity:**
   Through clustering, TPC retains frames with high semantic diversity, ensuring that pruned frames still capture complete information.
3. **Computational Efficiency:**
   By reducing the number of redundant frames, TPC significantly lowers the computational cost of Transformers. For instance, on MotionBERT, TPC reduces FLOPs by 51.8%.

Main approach

The structure of MHFormer with TPC



Evaluation Metric

MPJPE (Mean Per Joint Position Error) is a widely used metric in 3D human pose estimation that quantifies the average error between the predicted and ground-truth 3D joint positions.

Formula: MPJPE = $\frac{1}{\mathrm{N}} \sum_{i=1}^{N} \left\| p_i - \widehat{p}_i \right\|_2$

Where:

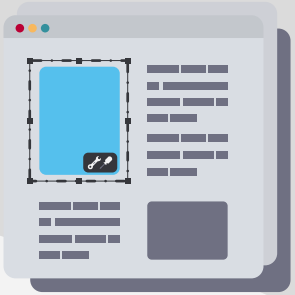
- N : The number of joints in the skeleton.
- p_i : The ground-truth 3D position of the i-th joint.
- $\widehat{p}_i$: The predicted 3D position of the i-th joint.
- $\| \|_2$: The Euclidean distance between the ground-truth and predicted positions.

A lower MPJPE value indicates higher accuracy, while a higher value means more difference from the ground truth.

Results & Analysis

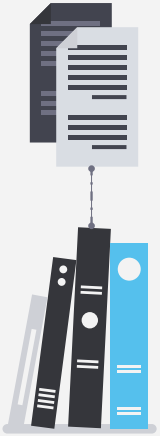
	Training time	MPJPE
MHformer	220 mins	43.87
MHformer with TPC	130 mins	44.14

Integrating TPC with MHformer reduces training time by 41% (220 mins to 130 mins) with only a slight increase in MPJPE (43.87mm to 44.14mm), offering an efficient trade-off between speed and accuracy.



Thanks for listening

Reference



- **MHFormer: Multi-Hypothesis Transformer for 3D Human Pose Estimation** : https://arxiv.org/pdf/2111.12707

- **Token Pruning Cluster** : https://arxiv.org/pdf/2311.12028

Contribution of each member

鍾名捷	25%	主題討論，環境架設，程式撰寫，簡報製作，報告。
洪子奇	25%	主題討論，論文研究，程式撰寫，簡報製作，報告。
李東諺	25%	主題討論，論文研究，程式撰寫，簡報製作，報告。
葉軒宇	25%	主題討論，環境架設，程式撰寫，簡報製作，報告。