3D Human Pose Estimation in Video

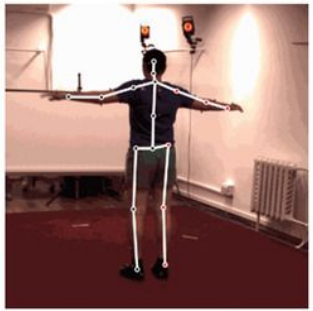
Alexandra Kissel, Bin Yang, Ganesh Arivoli

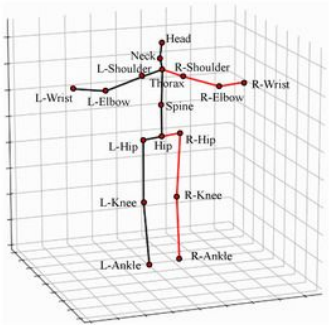
Problem

We are working on the problem of **estimating 3D human pose from video**. The input is a sequence of RGB frames, and the goal is to recover the 3D positions of human joints such as the head, shoulders, elbows, hips, knees, and ankles.

This is challenging because:

- A 2D image loses depth information
- Occlusions or clutter can hide body parts
- Movements can be fast or complex
- Models must generalize to new scenes and people





Our aim is to:
Compare transformer-based models with classical CNN models for 3D pose estimation in video

Why is it important

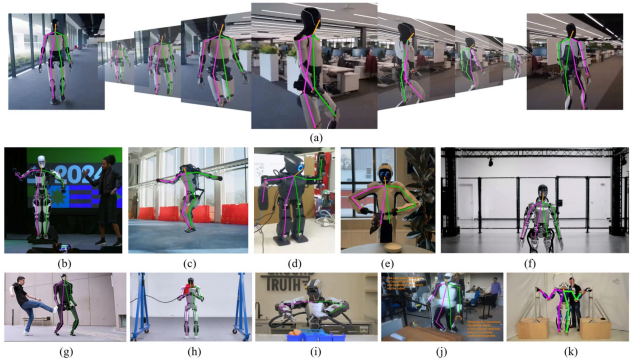
3D pose estimation is a cornerstone for many real-world systems:

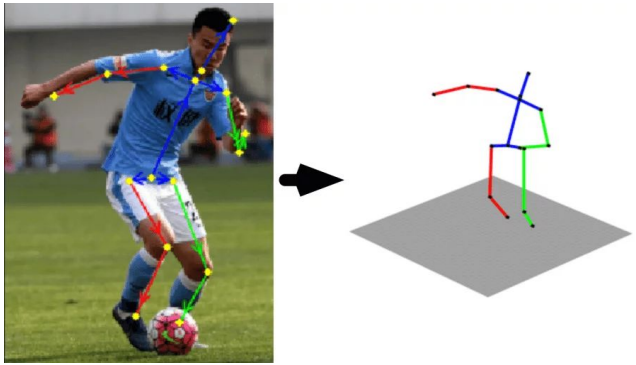
- **Robotics** – for safe human–robot interaction
- **AR/VR** – for tracking body movement in virtual environments
- **Healthcare & rehab** – analyzing patient motion
- **Sports analytics** – performance tracking and injury prevention

Current systems work well in clean conditions but struggle with:

- Complex motion sequences
- Depth uncertainty in single-camera video
- Hidden objects and clutter

Better 3D pose estimation increases **safety, robustness, and usability** across many practical applications.





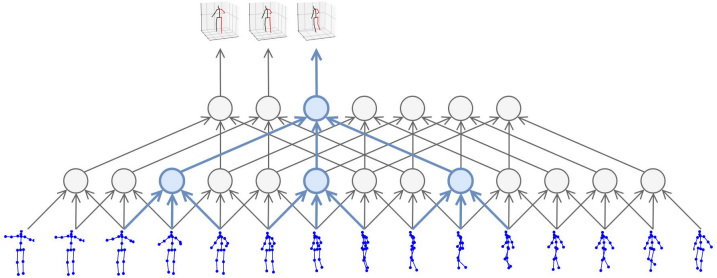
Our Method

We implemented and compared **two architectures** using identical 2D keypoints as input:

1. VideoPose3D (baseline CNN-based): Temporal fully connected network

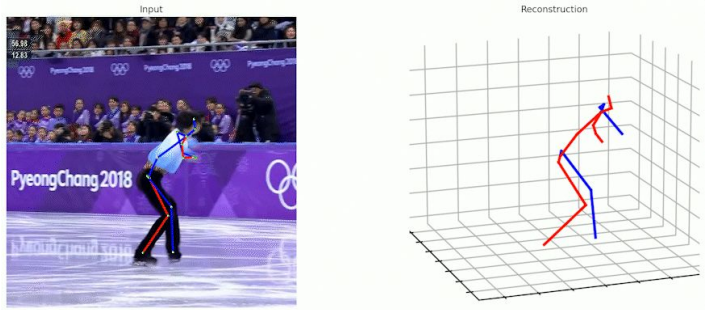
2. PoseFormerV2 (transformer-based): Uses self-attention to model global spatio-temporal relationships with frequency-domain representations.

Training dataset: Human 3.6M dataset (details in next slide)



We compare them using two methods:

1) Metrics using Human 3.6 dataset
2) Visual comparison using random videos:
   a) 2D Video frame to 2D keypoints: DetectronV2
   b) 2D key points uplifted to 3D coordinates

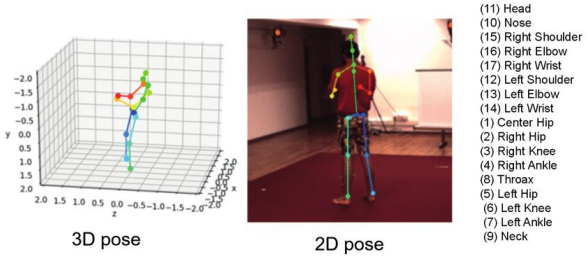


1) Pavllo, D., Feichtenhofer, C., Grangier, D., & Auli, M. (2019). *3D human pose estimation in video with temporal convolutions and semi-supervised training*
2) Zhao, Q., Zheng, C., Liu, M., Wang, P., & Chen, C. (2023). *PoseFormerV2: Exploring frequency domain for efficient and robust 3D human pose estimation*

About the Dataset and Metrics

Human3.6M contains 3.6 million video frames for 11 subjects
Each subject performs 15 actions that are recorded using four cameras

We adopt a 17-joint skeleton
Train on five subjects (S1, S5, S6, S7, S8)
Test on two subjects (S9 and S11).

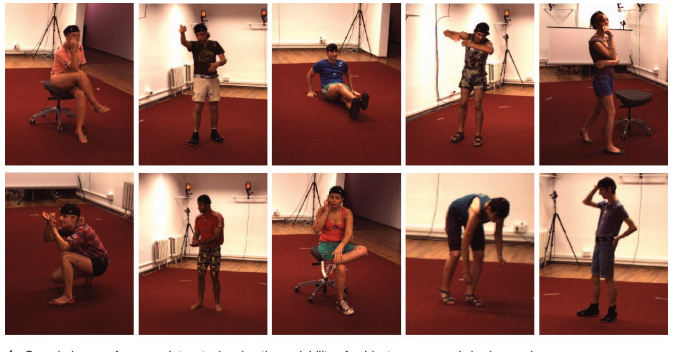


In our experiments, we consider three evaluation protocols:

1) P1 is the Mean per-joint position error (MPJPE) in millimeters which is the mean Euclidean distance between predicted joint positions and ground-truth joint positions.

2) P2 reports the error after alignment with the ground truth in translation, rotation, and scale (P-MPJPE - Procrustes-aligned Mean Per-Joint Position Error).

3) P3 aligns predicted poses with the ground-truth only in scale (N-MPJPE - Normalized Mean Per-Joint Position Error) for semi-supervised experiments.



Results and Comparison

S. No.	Error Metric	VideoPose3D	PoseFormerV2	% Improvement
1	MPJPE	46.8	45.2	3.4
2	P-MPJPE	36.5	35.6	2.5
3	N-MPJPE	45	43.8	2.7

All metrics measured in millimeters (mm)
• **Lower is better**
• PoseFormerV2 outperforms VideoPose3D across all metrics

Error Metrics:
• MPJPE (Protocol #1): Raw 3D position error
• P-MPJPE (Protocol #2): Error after pose alignment
• N-MPJPE (Protocol #3): Error after scale normalization

Results and Comparison

Sample video

Dancing:https://drive.google.com/file/d/14PDxucssJzmN9NkkmkfhxUfssWQiMDgt/view?usp=vids_web

Tennis:https://drive.google.com/file/d/1ReIH6KAJ9wUT-o9IMk6PIgngMb7MChxi/view?usp=vids_web

Running: https://drive.google.com/file/d/10z4IRNIBt8bZICjGdAfXEZRE4dcw0cvW/view?usp=vids_web

Output 3D coordinates comparison:

$$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}$$

	Mean 3D Error	Median	Max error
Tennis	1.067	1.059	1.717
Dancing	0.996	0.938	2.134
Running	1.009	0.897	1.765

Discussion

What we learned

- Transformers significantly outperform classical temporal models for 3D pose estimation.
- Data quality (2D keypoints) heavily influences 3D accuracy. (MediaPipe to Detectron)
- PoseFormer improves global joint understanding and adds stability when compared to the CNN based VideoPose3D.

Where this could go next

- Merge RGB images with **depth info** using LIDAR.
- Extend to multi-person pose estimation.
- Move toward real-time, on-device inference.

Overall, transformer-based architectures represent a strong next step for building robust and more generalizable 3D pose estimation systems.