Human Motion (or Action)

Qihang Fang

Introduction

- Understanding and modeling human motion is crucial for various applications.
- It enables advancements in robotics, healthcare, virtual reality (VR), sports analysis, and human-computer interaction.

Importance and Challenges

- Human motion analysis is complex due to the variability and intricacy of movement
- Research focuses on capturing, analyzing, and generating movements in realistic, adaptable ways.
- Key challenges
  - Complex skeletal structures
  - Variability in movement patterns
  - Integration with real-world environments (e.g., dynamic scenes)

Key Research Directions

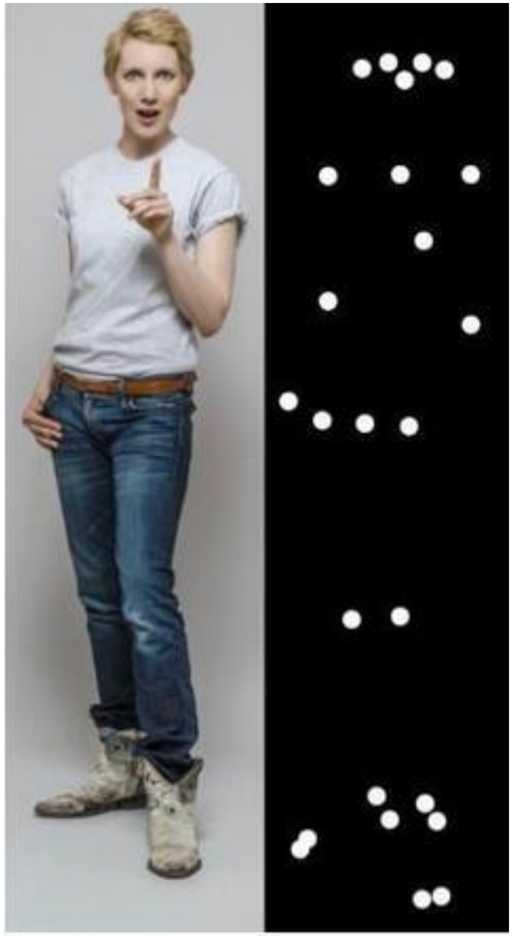
- This presentation covers major aspects of human motion modeling:
  - Modeling: Defining skeletal structures and kinematic models.
  - Capture: Motion capture technologies and 3D motion estimation.
  - Reconstruction: Techniques to reconstruct human models with clothing and skin.
  - Analysis: Methods for analyzing and classifying movements.
  - Generation: Motion synthesis for realistic animations and simulations.

Basic Model

Motion/action can be seen as a sequences of poses.

The question is:
How to represent 3D motion or pose sequences?

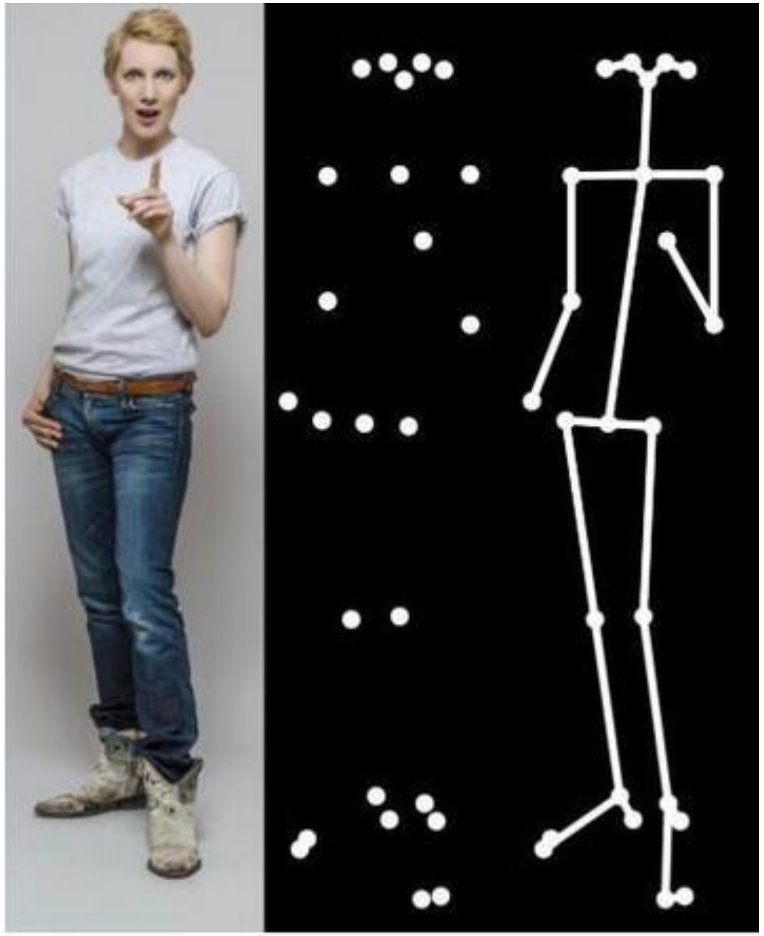
Basic Model – Key points



Human motion models identify specific key points or joints to simplify motion tracking and analysis.

Key points typically include hands, shoulders, elbows, hips, knees, and ankles, representing critical movement locations.

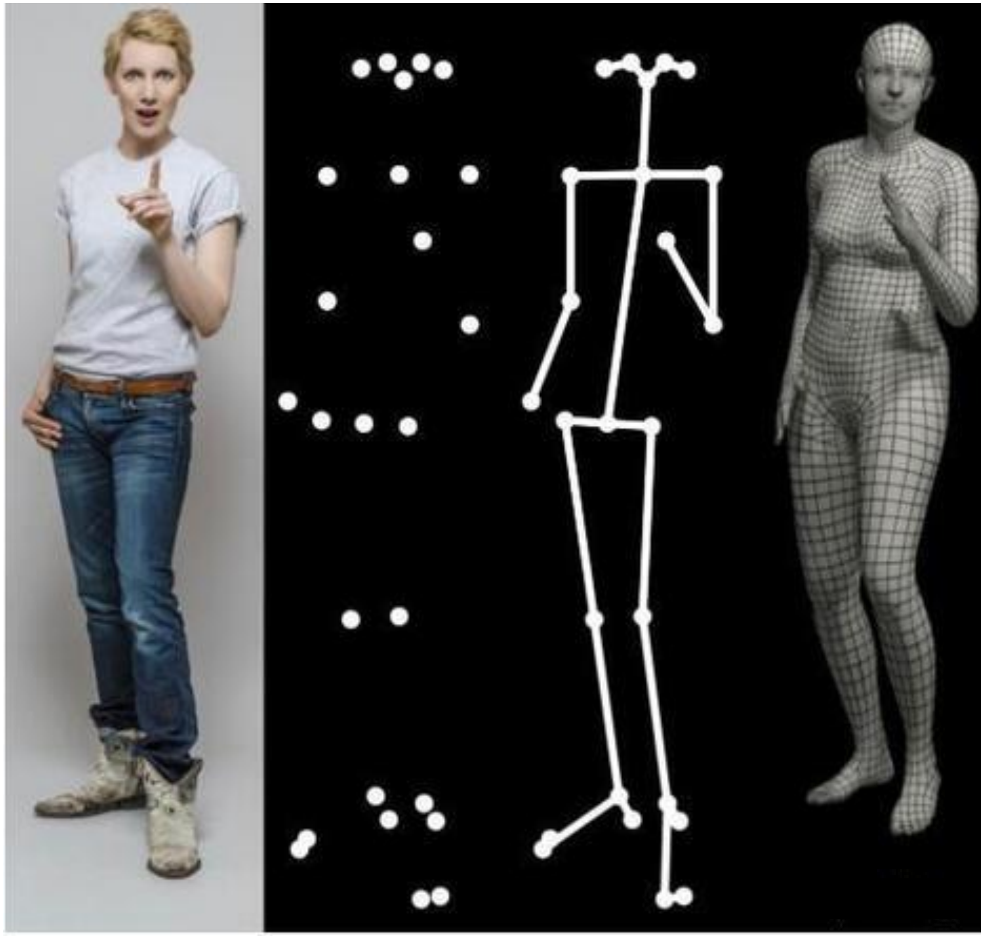
Basic Model – Kinematic Tree



Human motion models identify specific key points or joints to simplify motion tracking and analysis.

The kinematic tree is widely used in robotics and animation for its ability to replicate natural, constrained movement.

Basic model – SMPL Model

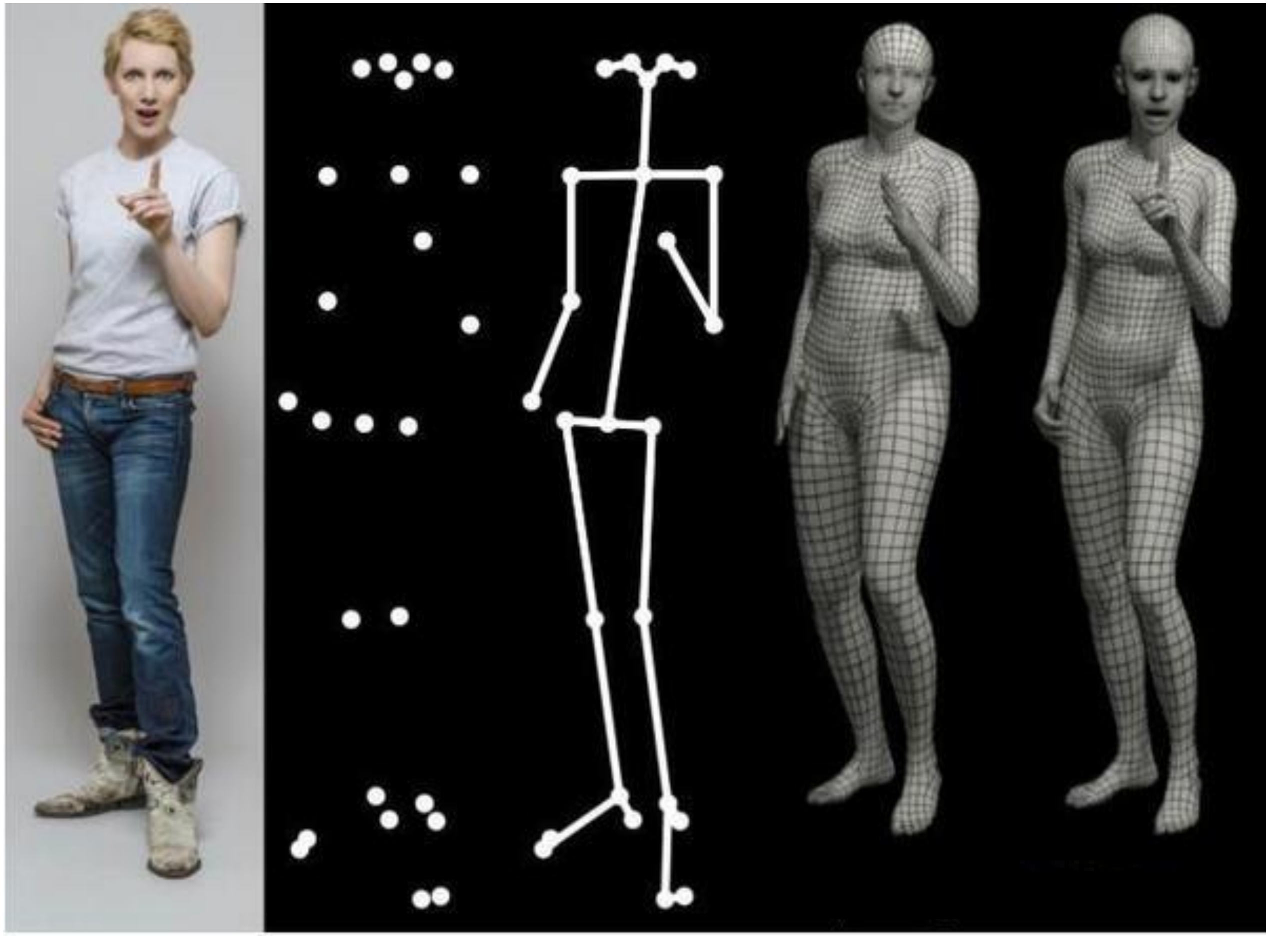


The SMPL model is a 3D parametric model representing human body shape and pose in high fidelity.

SMPL is widely used for its ability to recreate detailed body geometry while maintaining efficient computation.

Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., & Black, M. J. (2015). SMPL: A Skinned Multi-Person Linear Model. ACM Transactions on Graphics (TOG), 34(6), 1-16.

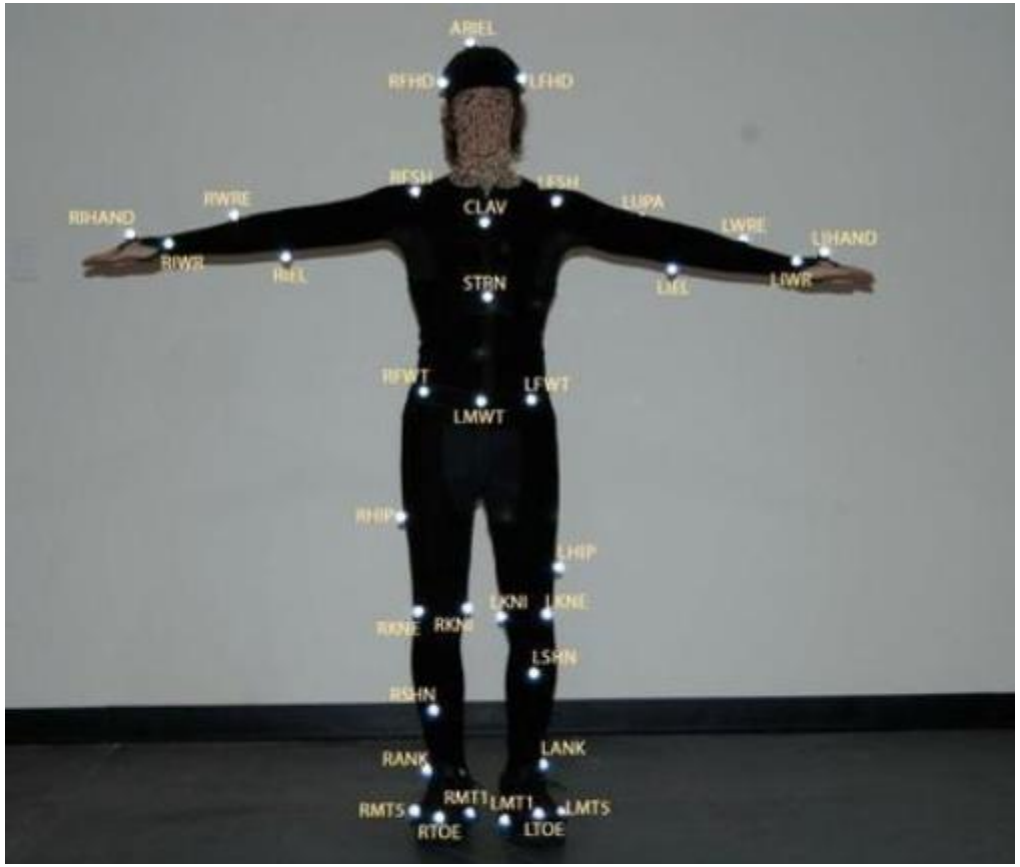
Basic model



SMPL-X extends the SMPL model to include facial expressions and detailed hand movements, adding greater versatility.

Pavlakos, G., Grigorev, S., Choutas, V., Bolkart, T., Osman, A. A. A., Tzionas, D., & Black, M. J. (2019). Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 10975-10985.

Introduction to Motion Capture



Motion Capture (MoCap) is a technique for digitally recording human movements and translating them into computer models.

Strength:

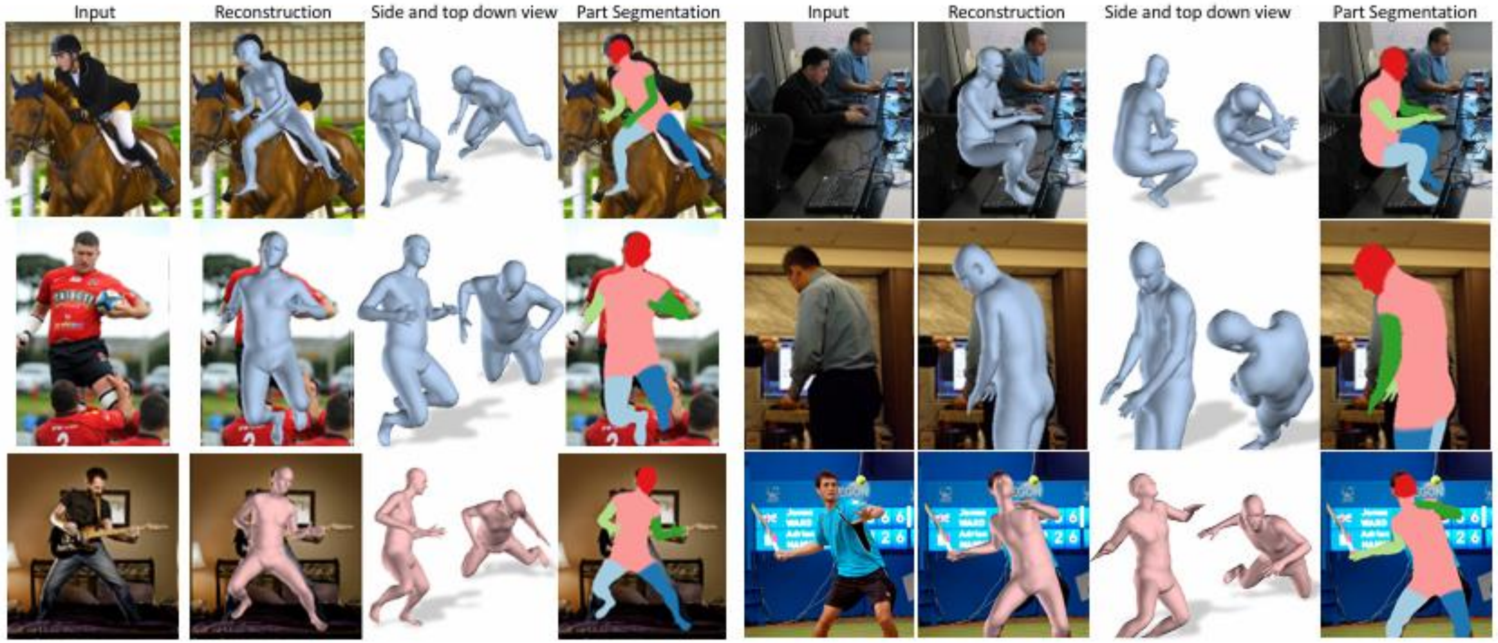
- Precise

Drawback:

- Expensive
- Limited scene

Motion estimation

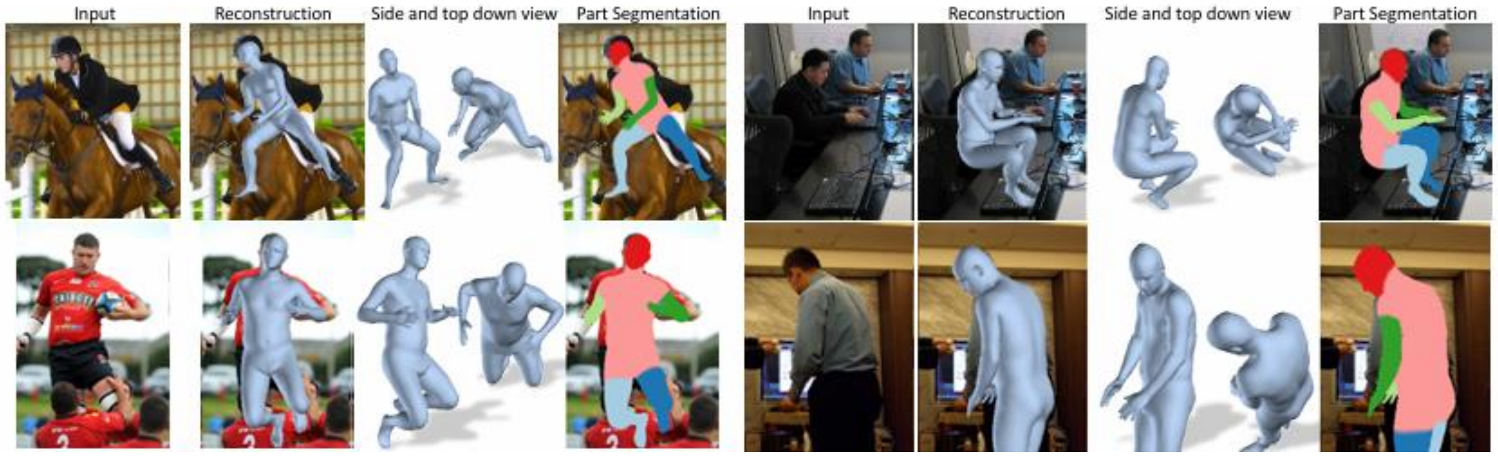
- 3D Human Motion Estimation involves estimating body pose and motion in three dimensions based on input data, typically images or video.



Kanazawa, A., Black, M. J., Jacobs, D. W., & Malik, J. (2018). End-to-End Recovery of Human Shape and Pose. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 7122-7131.

Challenges: Occlusion

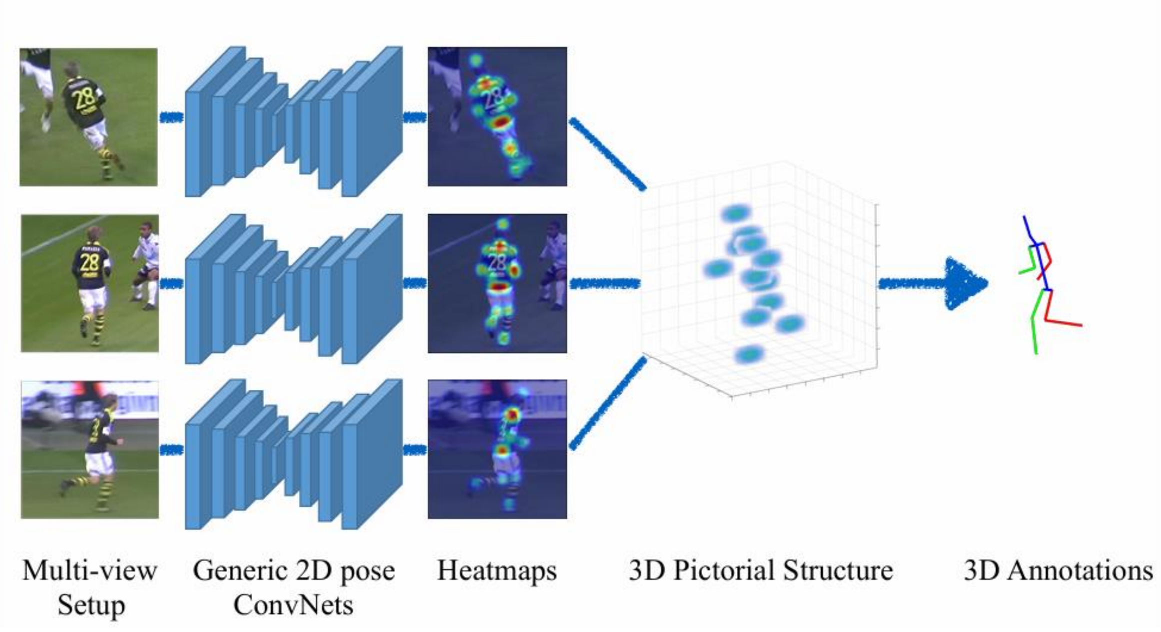
- Occlusion occurs when body parts are hidden from the camera, making it difficult to estimate joint positions accurately.
- This issue is prevalent in dynamic scenes and crowded environments.



Kanazawa, A., Black, M. J., Jacobs, D. W., & Malik, J. (2018). End-to-End Recovery of Human Shape and Pose. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 7122-7131.

Challenges: Occlusion

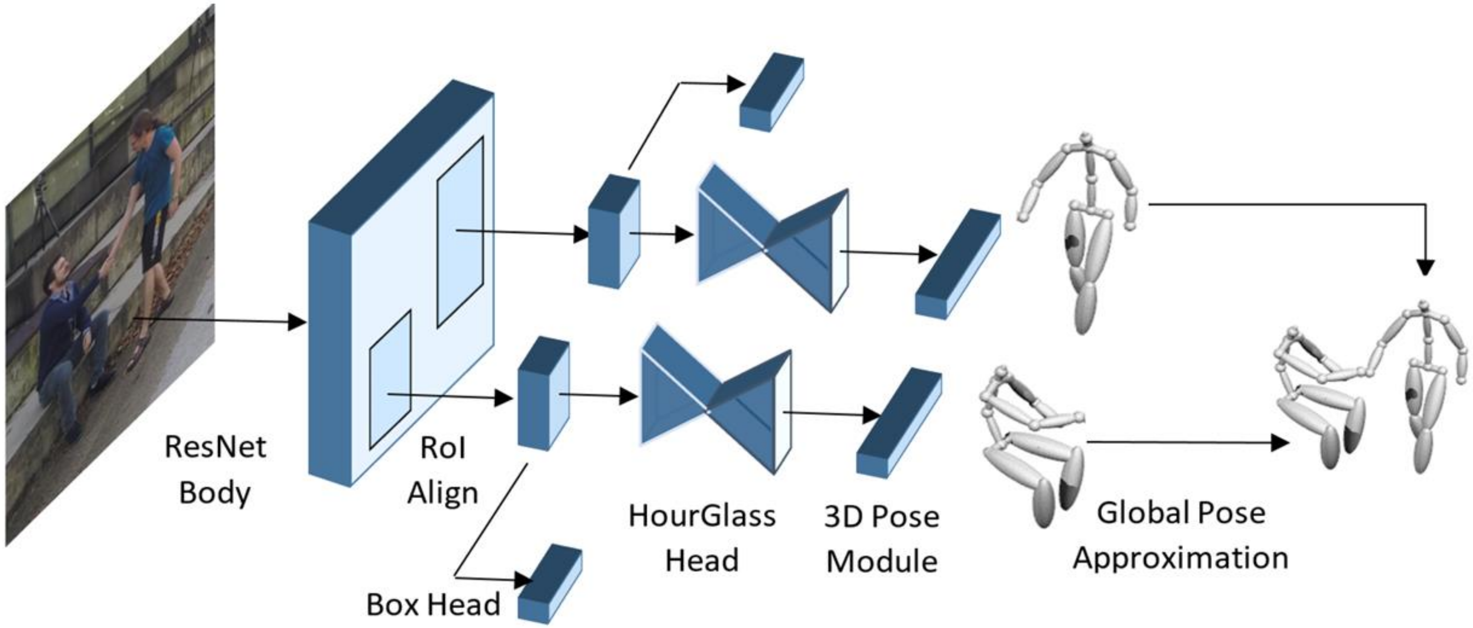
- Easiest solution:
  - Add more views



Pavlakos, G., Zhou, X., Chan, A., Derpanis, K. G., & Daniilidis, K. (2017). Harvesting Multiple Views for Marker-less 3D Human Pose Annotations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 6903-6911.

Challenges: Occlusion

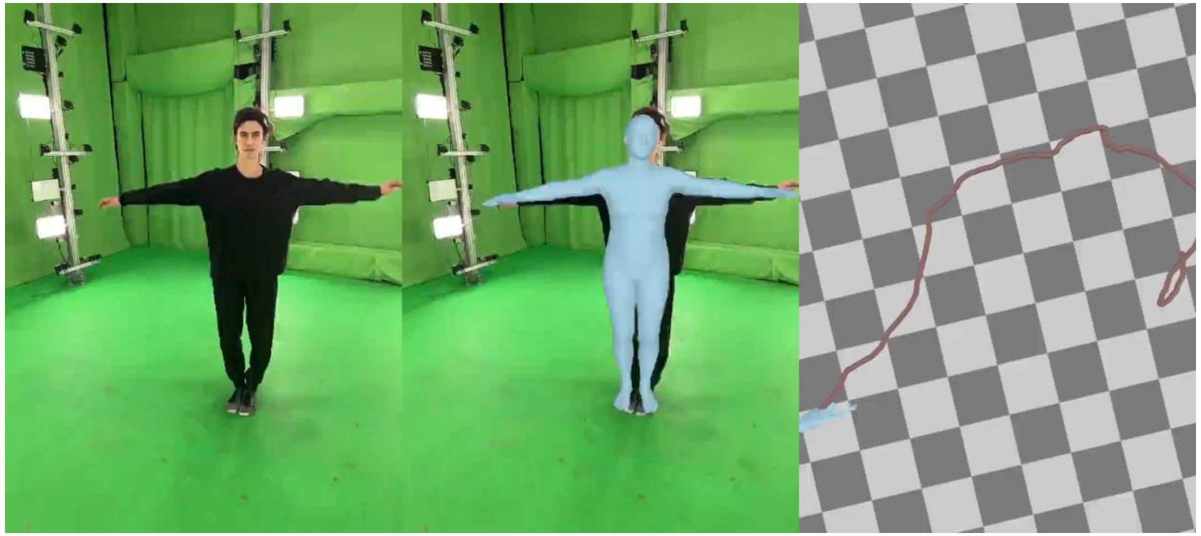
- Another solution:
  - Motion prior
    - For example: utilize interaction between different people



Dabral, R., Kusupati, U., Afaque, S., Mundhada, A., Sharma, A., & Jain, A. (2019). Multi-Person 3D Human Pose Estimation from Monocular Images. In Proceedings of the British Machine Vision Conference (BMVC).

Challenges: Dynamic Camera Movements

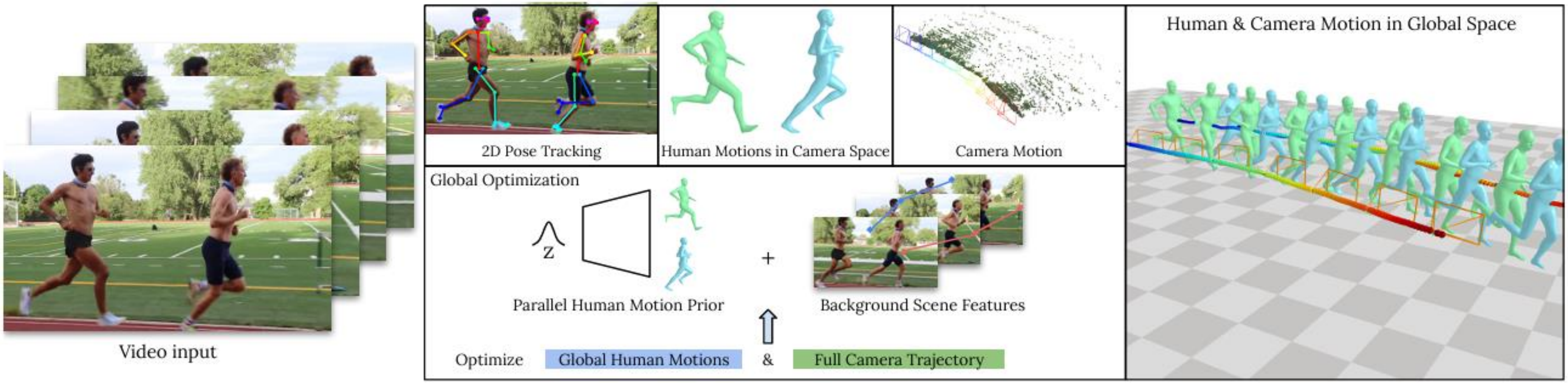
- Dynamic cameras introduce complexities due to changing viewpoints and backgrounds.
- Handling dynamic cameras requires robust algorithms that can stabilize motion estimation despite camera movements.



Shin, Soyong, et al. "Wham: Reconstructing world-grounded humans with accurate 3d motion." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

Challenges: Dynamic Camera Movements

- Popular solution:
  - Estimation with SLAM
    - Drawback： slow, low accuracy of



Shin, Soyong, et al. "Wham: Reconstructing world-grounded humans with accurate 3d motion." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

Challenges: Depth Prediction

- Monocular (single-camera) systems lack depth information, complicating accurate 3D pose estimation. Methods for addressing depth ambiguity include monocular depth prediction algorithms or incorporating scene context.
- Solution: Depth prediction

Human Shape Reconstruction

- Human shape reconstruction focuses on creating accurate 3D models of the human body from 2D images or 3D scans.
- The task aims to capture details like body shape, posture, and, in some cases, clothing, to generate lifelike digital humans.
- Applications range from VR/AR, film and gaming industries, to healthcare and biometric analysis.

Huang, Yangyi, et al. "Tech: Text-guided reconstruction of lifelike clothed humans." 2024 International Conference on 3D Vision (3DV). IEEE, 2024.

Action Recognition in Video

- Action recognition involves identifying specific actions from video sequences, like 'running,' 'jumping,' or 'dancing.
  - Crucial for applications in sports, surveillance, and human-computer interaction.
  - Inputs: Videos or motion sequences

Soomro, K. "UCF101: A dataset of 101 human actions classes from videos in the wild." arXiv preprint arXiv:1212.0402 (2012).

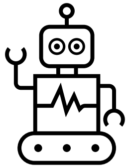
Challenges – Complex Backgrounds

- Crowded environments and rapidly changing backgrounds add complexity to action recognition.

Zhao, Wang, et al. "Particlesfm: Exploiting dense point trajectories for localizing moving cameras in the wild." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022.

Challenges – Complex Backgrounds

- Crowded environments and rapidly changing backgrounds add complexity to action recognition.
- Bias in the dataset environment can also affect the judgment of action type.

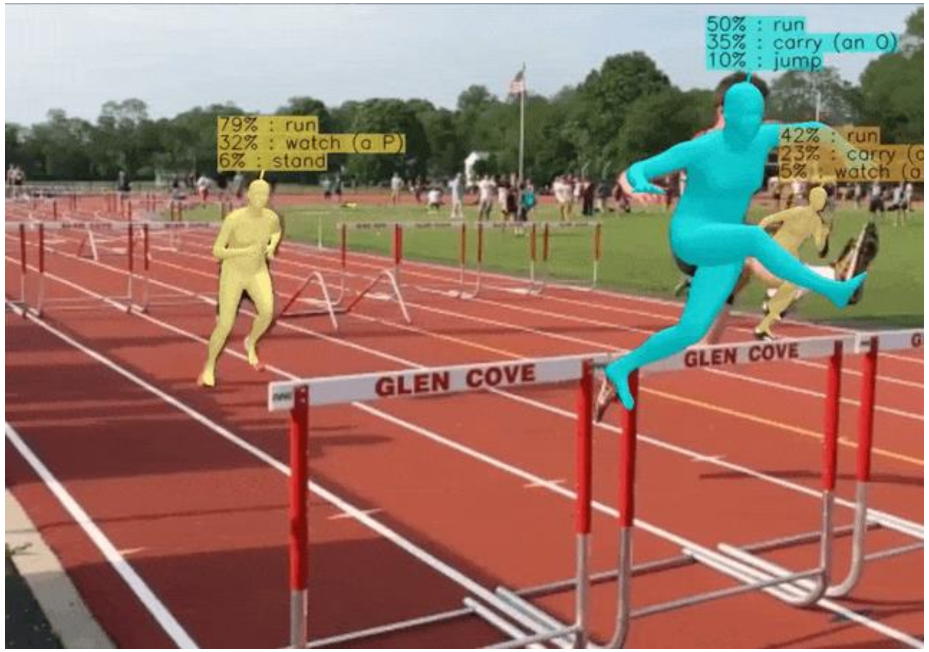


I see a stadium.

Thus, the action is running.

Challenges – Complex Backgrounds

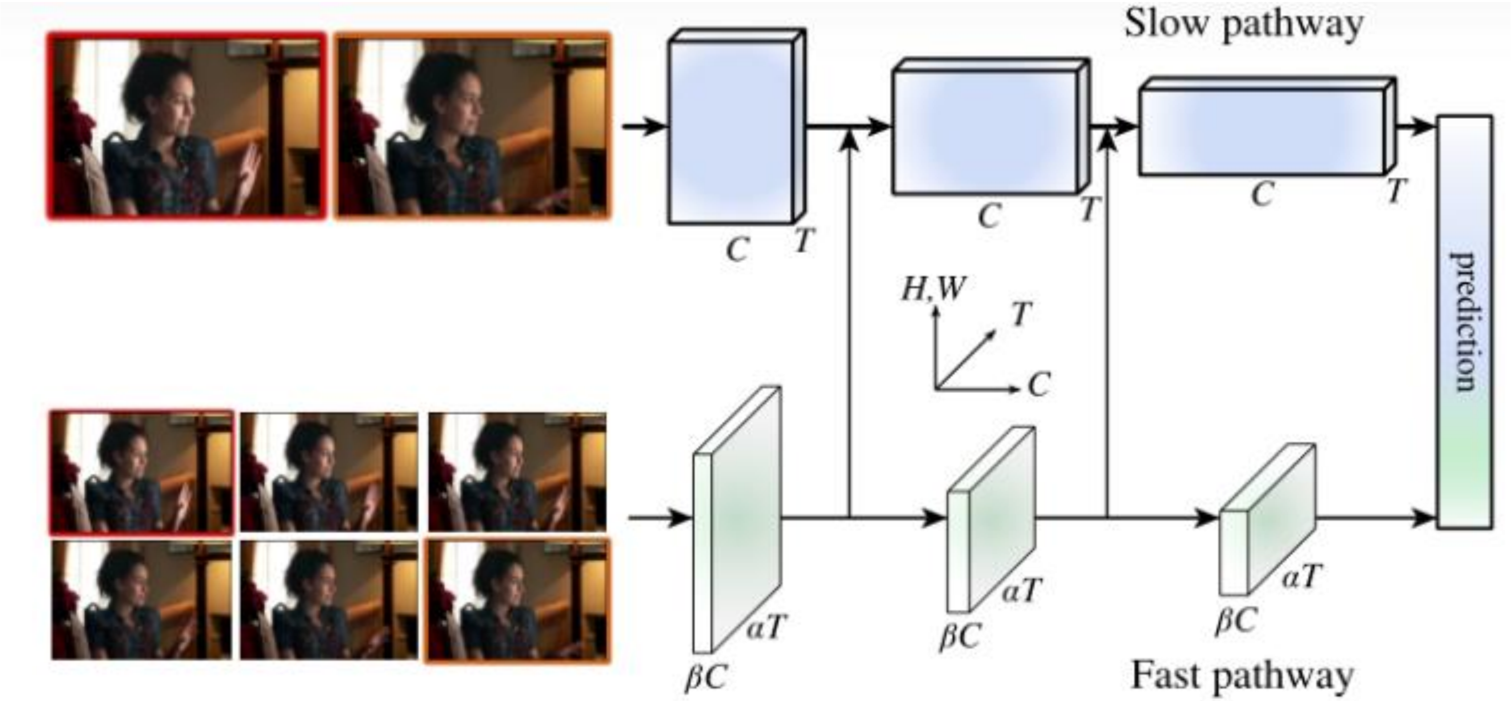
- Solution:
  - New dataset -- Expensive
  - Pay more attention on the motion



Rajasegaran, Jathushan, et al. "On the benefits of 3d pose and tracking for human action recognition." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

Challenges – Intra-Class Variability

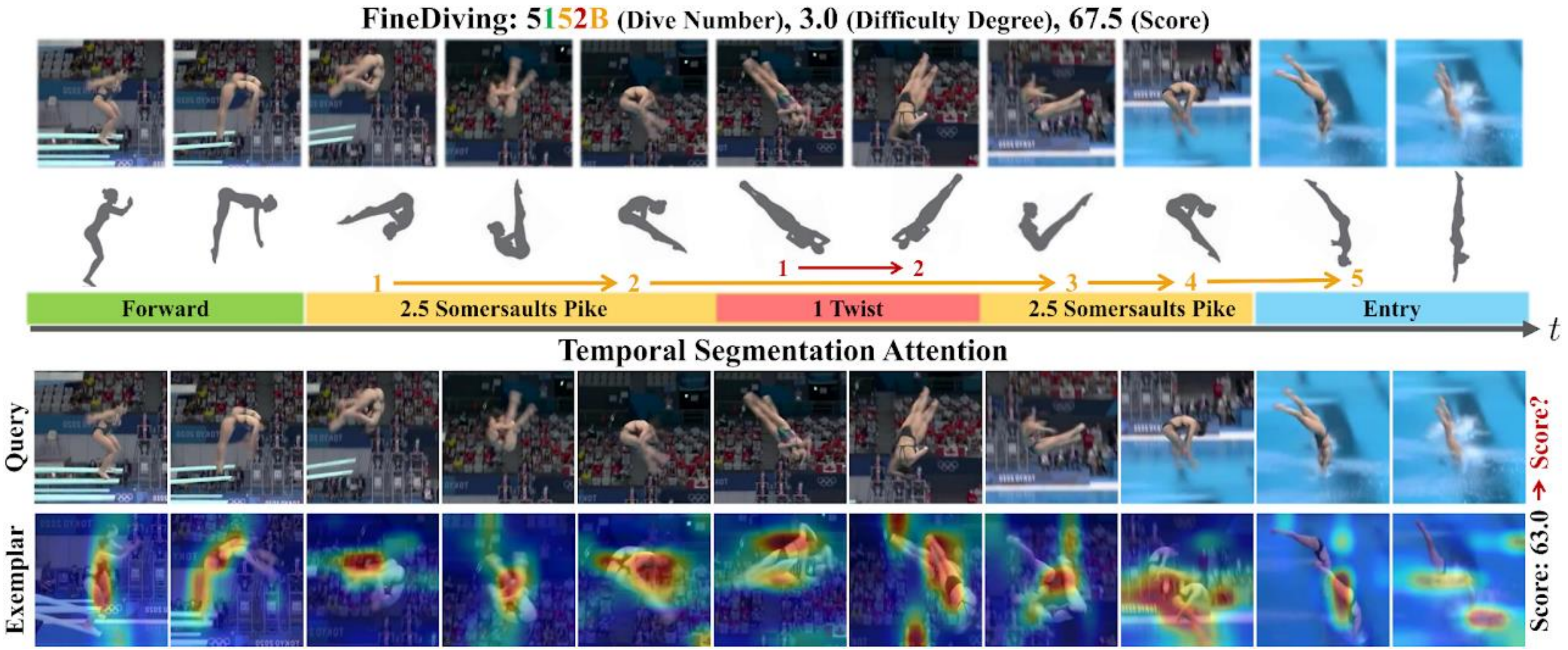
- The same action can appear different depending on the performer, their posture, their speed or the environment.

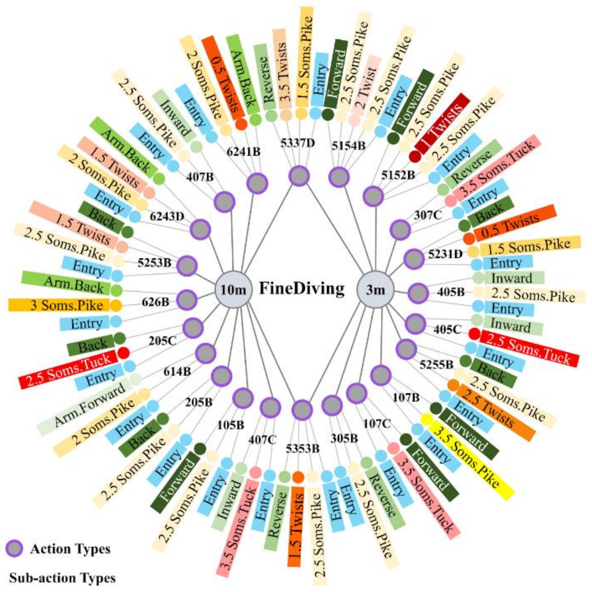


Feichtenhofer, Christoph, et al. "Slowfast networks for video recognition." Proceedings of the IEEE/CVF international conference on computer vision. 2019.

Challenges – Fine-Grained

- Fine-grained actions require capturing subtle differences in motion and object interaction.

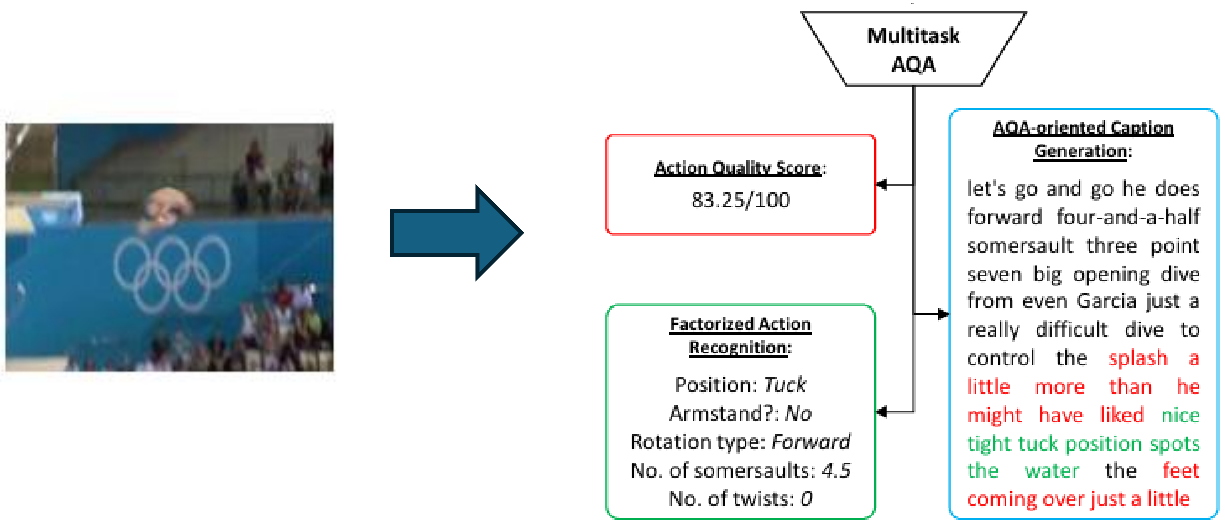




Xu, Jinglin, et al. "Finediving: A fine-grained dataset for procedure-aware action quality assessment." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

Action quality Assessment

- Action Quality Assessment (AQA) evaluates the quality of an action, rather than merely identifying it.
- Common in areas like sports, physical therapy, and skill evaluation, AQA quantifies aspects like form, smoothness, and efficiency.



Parmar, Paritosh, and Brendan Tran Morris. "What and how well you performed? a multitask learning approach to action quality assessment." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.

Video Understanding

- Video understanding encompasses not only recognizing actions but also capturing scene context, interactions, and activities.
- Aims to interpret complex scenarios in video data, like social interactions or multi-agent activities.

Mangalam, Karttikeya, Raiymbek Akshulakov, and Jitendra Malik. "Egoschema: A diagnostic benchmark for very long-form video language understanding." Advances in Neural Information Processing Systems 36 (2023): 46212-46244.

Introduction to Motion Generation

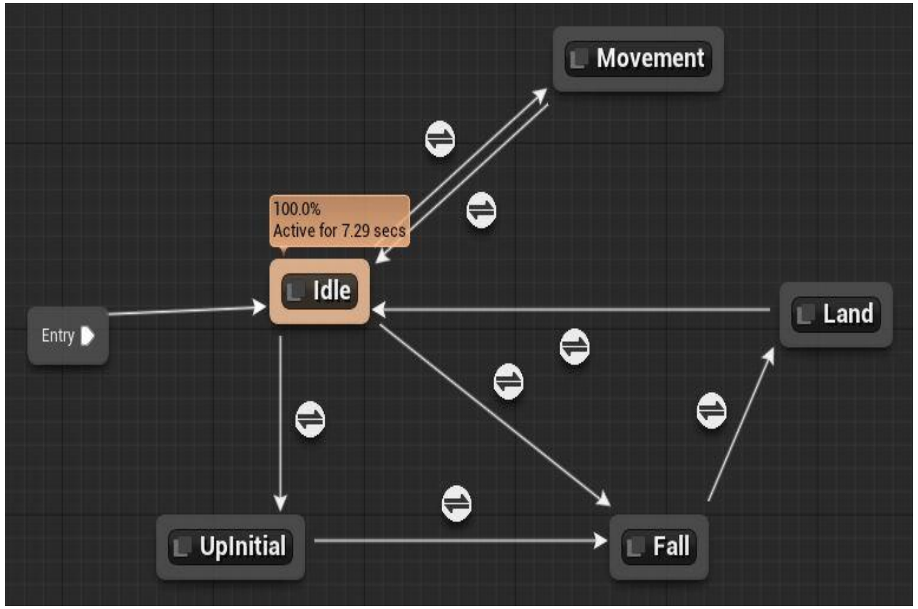
- Motion generation creates realistic human movements for applications in animation, VR, gaming, and robotics.

Games

Animations

State Machine-Based Motion Generation

- **Advantage**:
  - Simple and proven
  - Low memory usage
  - Fast
- **Disadvantage**:
  - Unrealistic transitions between different types of motions
  - Labor-consuming



Motion Matching

- Motion matching selects pre-recorded motion clips based on similarity to desired poses, enabling seamless transitions.
- Widely used in gaming to create fluid animations that adapt to player input.

Motion Matching

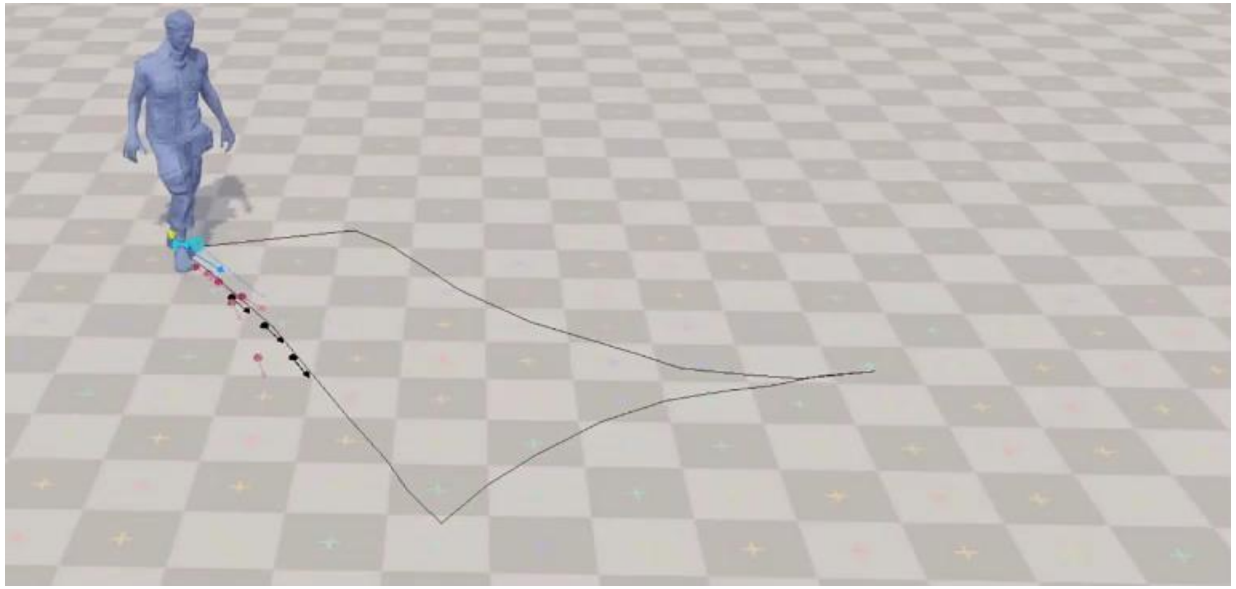
- Motion matching selects pre-recorded motion clips based on similarity to desired poses, enabling seamless transitions.
- Widely used in gaming to create fluid animations that adapt to player input.

- At first, we need a motion capture dataset.

Motion Matching

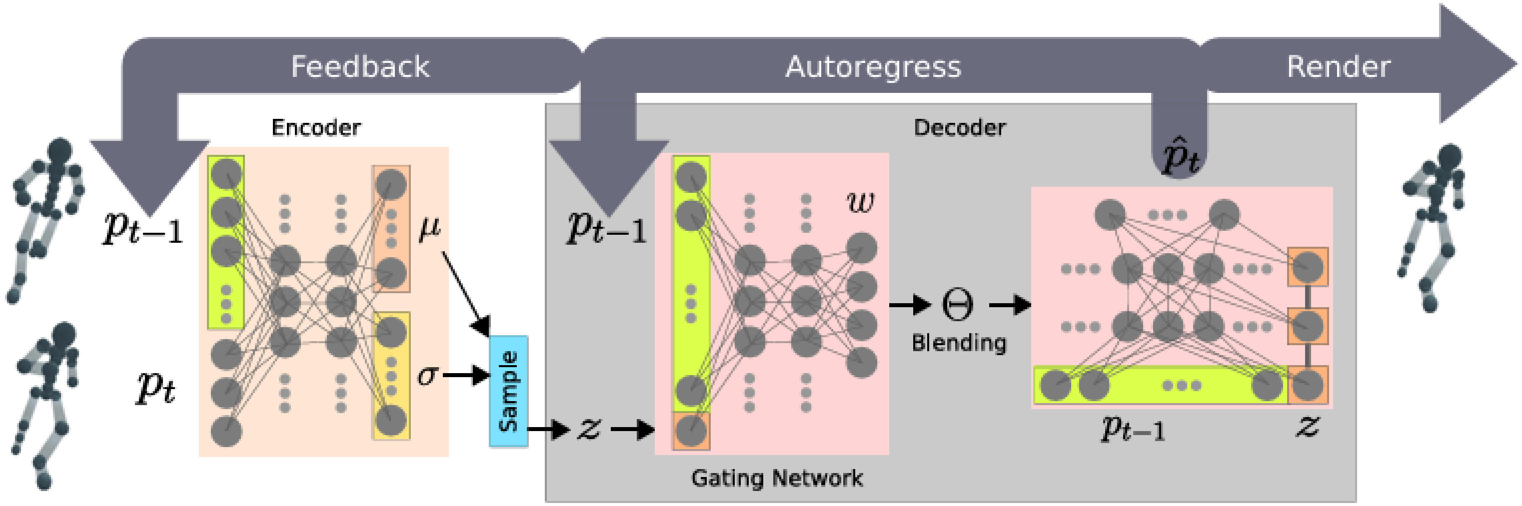
- Motion matching selects pre-recorded motion clips based on similarity to desired poses, enabling seamless transitions.
- Widely used in gaming to create fluid animations that adapt to player input.

- At first, we need a motion capture dataset.
- Then, we select the next frame from the dataset by pre-defined rules.



Data-driven Motion Generation

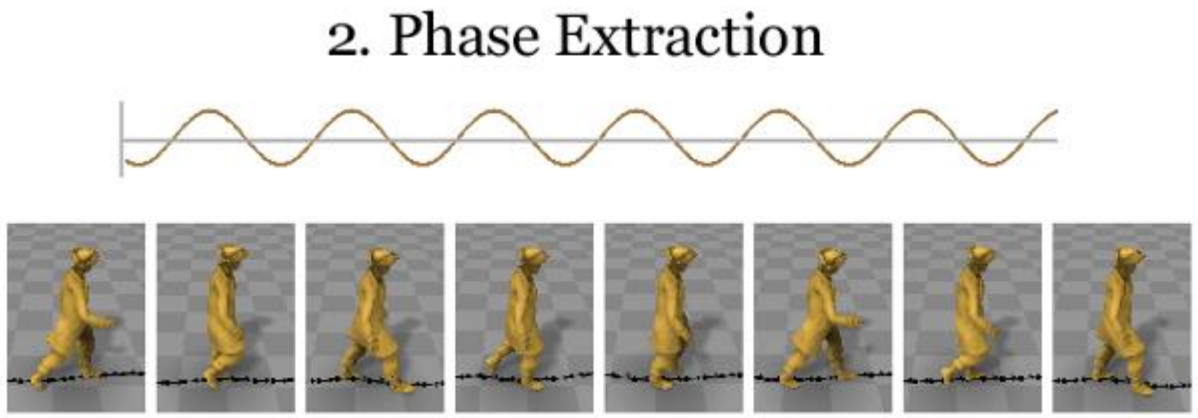
- Data-driven motion generation leverages large datasets of recorded human motions to learn natural movement patterns.
- This approach enables models to synthesize realistic, data-informed motions by extracting patterns directly from real-world data.

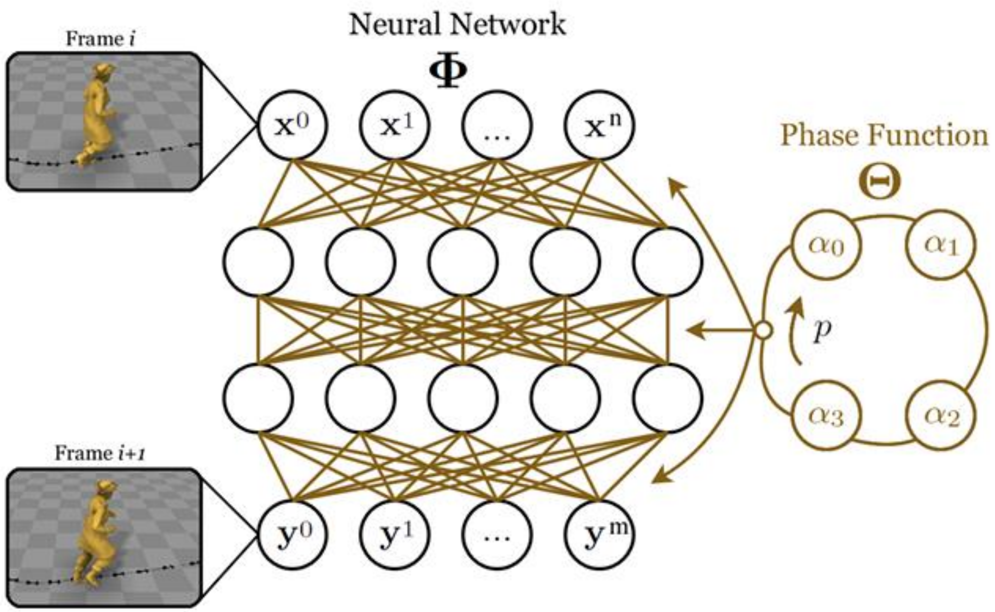


Ling, Hung Yu, et al. "Character controllers using motion vaes." ACM Transactions on Graphics (TOG) 39.4 (2020): 40-1.

Phase-Based Motion Generation

- Phase-based approaches use a continuous phase variable to control motion, allowing smooth transitions and timing adaptation.



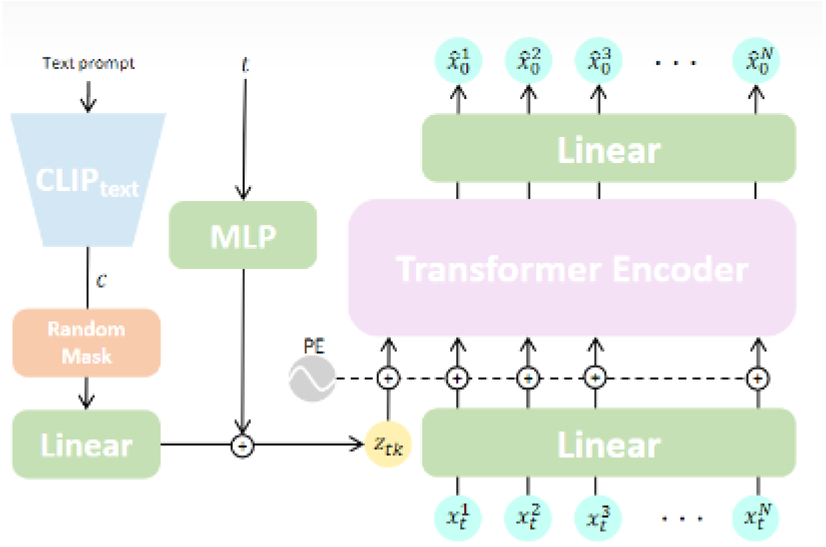


Holden, Daniel, Taku Komura, and Jun Saito. "Phase-functioned neural networks for character control." ACM Transactions on Graphics (TOG) 36.4 (2017): 1-13.

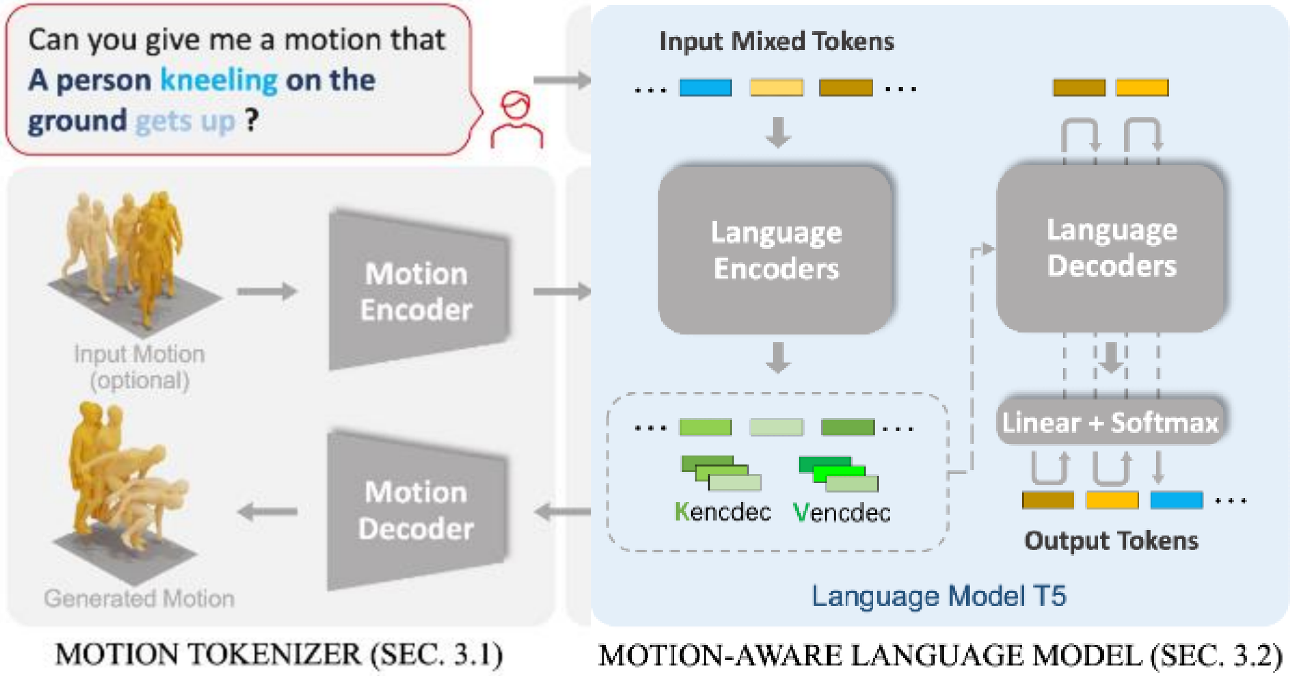
Phase-Based Motion Generation

Starke, Sebastian, et al. "Local motion phases for learning multi-contact character movements." ACM Transactions on Graphics (TOG) 39.4 (2020): 54-1.

New Era – Text-based Motion Generation



[MDM]



[MotionGPT]

Thanks