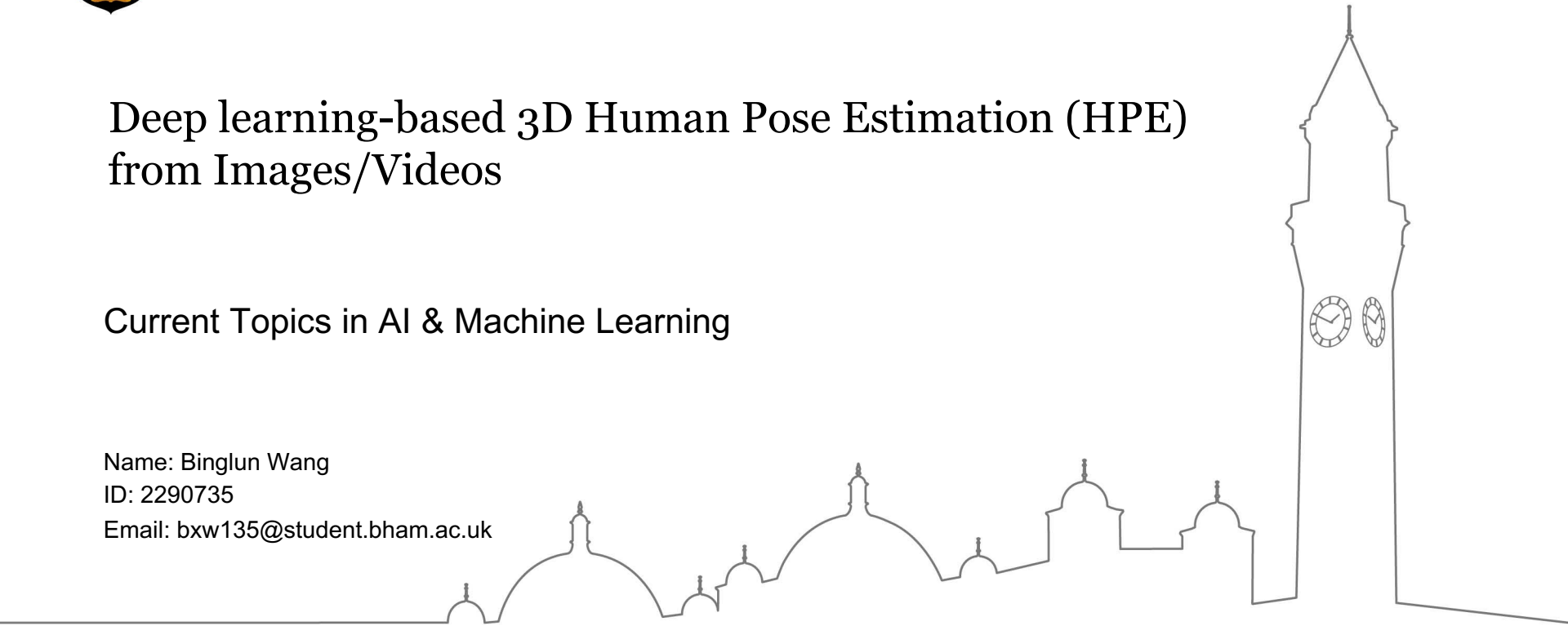
UNIVERSITY OF BIRMINGHAM

# Deep learning-based 3D Human Pose Estimation (HPE) from Images/Videos

## Current Topics in AI & Machine Learning

Name: Binglun Wang
ID: 2290735
Email: bxw135@student.bham.ac.uk



Highlight

- 3D Human Pose Estimation (**HPE**)

Deep Learning **(DL)** Method

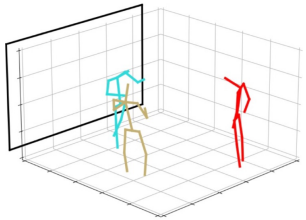
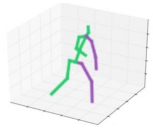


Fig 1. Example of 3D HPE [1]

- Human Body Model:


Fig 2. *Left:* Skeleton [2], *Middle:* SMPL [3], *Right:* Surface[4]

- Categories :



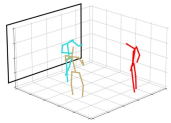


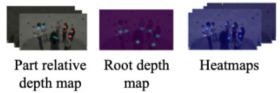



Fig 3. Categories [1][5][27]

- Dataset

Human3.6M [6],
3DPW [7],
MPI-INF-3DHP2 [2],
DensePose-COCO [4],
etc.

- Evaluation Metrics




$$MPJPE = \frac{1}{N}\sum_{i=1}^{N} \|J_i - J_i^*\|_2,$$

Where $N$ is the number of joints, $J_i$ and $J_i^*$ are the ground truth position and the estimated position of the $i_{th}$ .

PMPJPE,
NMPJPE,
MPVE [8],
etc.


[1] “Distribution-Aware Single-Stage Models for Multi-Person 3D Pose Estimation.”, [Wang et al., CVPR2022]
[2] “Monocular 3d human pose estimation in the wild using improved CNN supervision.”, [Mehta et al., 3DV2017]
[3] “SMPL: A skinned multi-person linear model.” [Loper et al., TOG 34.6(2015): 1-16.]
[4] “Densepose: Dense human pose estimation in the wild.” [Güler et al., CVPR2018]
[5] “Deep learning-based human pose estimation: A survey.” [Zheng et al., Tsinghua Science and Technology, 2019]
[6] “Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments.” [Ionescu et al., IEEE TPAMI2014]
[7] "Recovering accurate 3d human pose in the wild using imus and a moving camera." [Marcard et al., ECCV2018]
[8] “Learning to Estimate 3D Human Pose and Shape from a Single Color Image.” [Pavlakos et al., CVPR2018]
[27] “SMAP: Single-Shot Multi-Person Absolute 3D Pose Estimation.” [Zhen et al., ECCV2020]

- Action correction and online coaching.
- Clothes parsing
- AR/VR

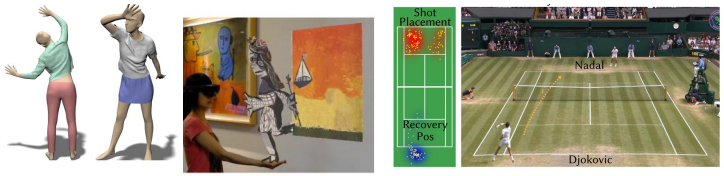



Fig 4. *Left:* Clothes parsing [10], *Middle:* AR [11], *Right:* VR [12]

- Action recognition, prediction, detection

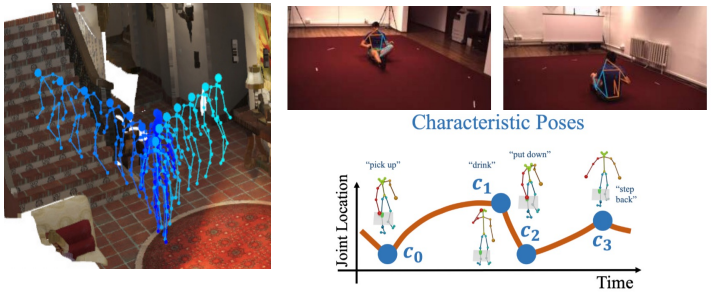
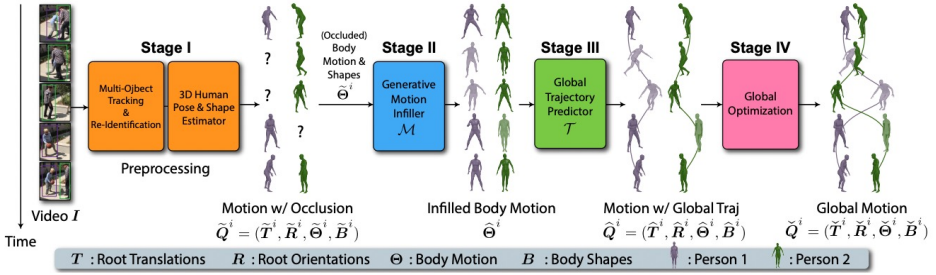



Fig 5. Examples. *Left:* [13], *Right-top:* [14], *Right-bottom:* [15]

- Action recognition, prediction, detection (cont.)



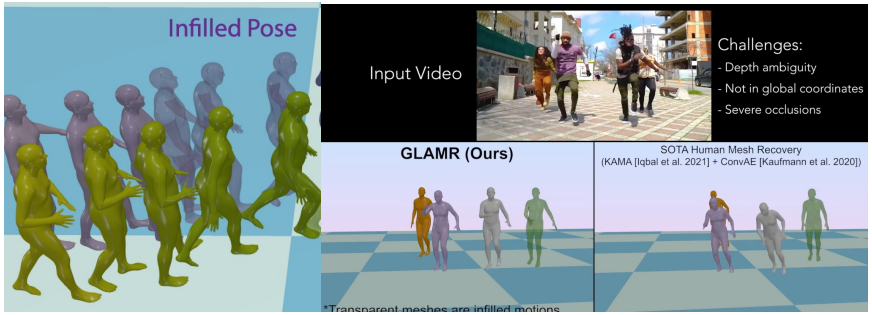




Fig 6. Examples. GLAMR [16]

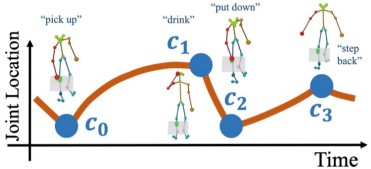
[9] “AI Coach: Deep Human Pose Estimation and Analysis for Personalized Athletic Training Assistance.” [Wang et al., ACM MM2019]
[10] “TailorNet: Predicting Clothing in 3D as a Function of Human Pose, Shape and Garment Style.” [Patel et al., CVPR2020]
[11] “Photo Wake-Up: 3D Character Animation From a Single Photo.” [Weng et al., CVPR2019]
[12] "Vid2player: Controllable video sprites that behave and appear like professional tennis players." [Zhang et al., TOG 40.3 (2021): 1-16.]
[13] “Long-term human motion prediction with scene context.” [Cao et al., ECCV2020(Oral)]
[14] “View-Invariant Probabilistic Embedding for Human Pose.” [Sun et al., ECCV2020]
[15] “Forecasting Characteristic 3D Poses of Human Actions.” [Diller et al., CVPR2022]
[16] "GLAMR: Global Occlusion-Aware Human Mesh Recovery with Dynamic Cameras." [Yuan et al., CVPR2022(Oral)]

- In 3DHPE, what are possibilities of interaction between human and computer & scenes?
  - Robot [17]

  - GAME [28]


  - 3D scenes [15]

Characteristic Poses




  - Metaverse [18]

Fig 7. Pictures in Interaction

- Apart from the main methods, direct collection dataset, what are the alternative methods as supplements?
  - From 2D data [19][20][21]

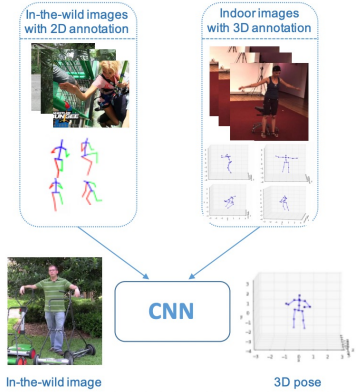



Fig 8. A weakly supervised method[19]

  - From Game[13][22]

Fig 9. Example from GTA-IM[13]

  - Domain Adaption
  - Data Augmentation
  - Others

[17] Picture from “STRANDS” project , Intelligent Robotics Lab (IRLab), University of Birmingham
[18] Picture from Ready Player One (film), Directed by Steven Spielberg
[19] “Towards 3D Human Pose Estimation in the Wild: A Weakly-Supervised Approach.” [Zhou et al., ICCV2017]
[20] "In the Wild Human Pose Estimation Using Explicit 2D Features and Intermediate 3D Representations.”[Habibie et al., CVPR2019]
[21] “Unsupervised 3D Pose Estimation With Geometric Self-Supervision.” [Chen et al., CVPR2019]
[22] “Learning from synthetic humans.” [Varol et al., CVPR2017]
[28] “VRChat.” [VRChat Inc. , https://hello.vrchat.com/]

- Directly from images
  (Images: the projection of 3D to 2D.)
- Background variation, camera movement, fast move, illumination changes, etc.
- Occlusions.

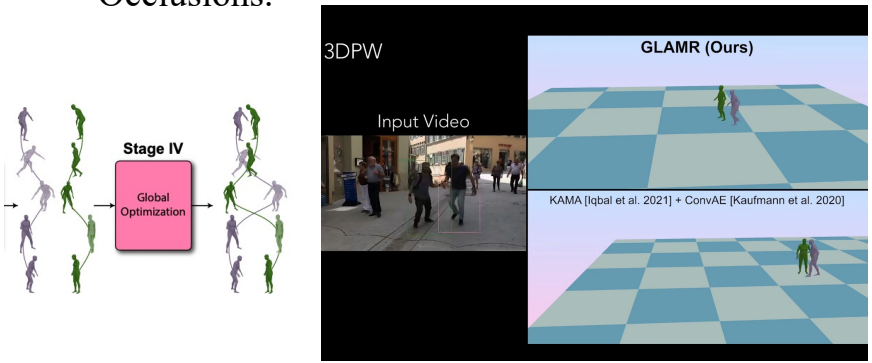



Fig 10. camera movement challenge [16]

- How to fuse information from multiple cameras.
- In-the-Wild Scenario
- Reduce the number of parameters while preserving quality

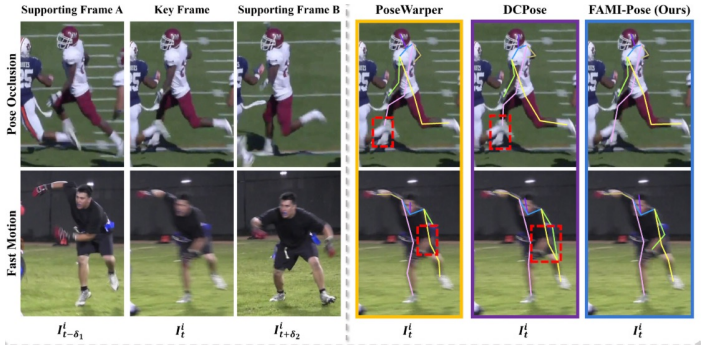



Fig 11. Occlusions and fast move challenge [23]

- Neural Architecture Search in 3D HPE [24][25][26]

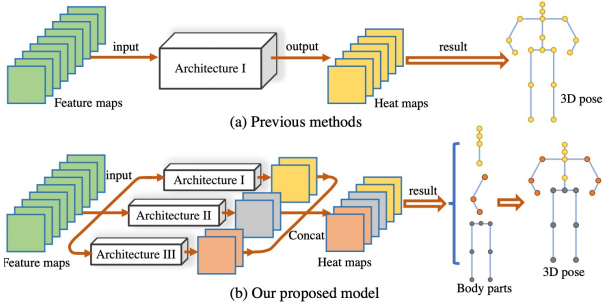



Fig 12. A Neural Architecture Search method [25]

- 3D HPE shape reconstruction from videos are not smooth and continuous.

[23] “Temporal Feature Alignment and Mutual Information Maximization for Video-Based Human Pose Estimation” [Liu et al., CVPR2022(Oral)]
[24] “Neural Architecture Search: A Survey.” [Elsken et al., JMLR2019]
[25] “Towards Part-aware Monocular 3D Human Pose Estimation: An Architecture Search Approach.” [Chen et al., ECCV2020]
[26] “EfficientPose: Efficient Human Pose Estimation with Neural Architecture Search.” [Zhang et al., *Computational Visual Media* 7.3 (2021): 335-347.]