Monocular Dynamic Motion Capture: A Regression-Optimization Hybrid Approach

Supervisor: Hedvig Kjellström

Athanasios (Thanos) Charisoudis
thacha@kth.se

M.Sc. in Machine Learning
KTH Royal Institute of Technology

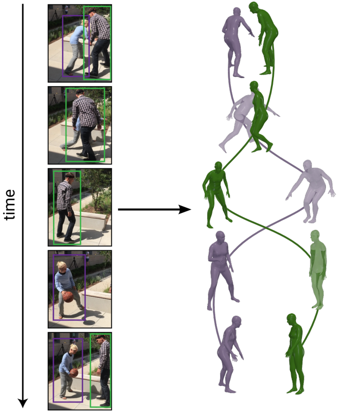
April 19, 2024

Monocular Dynamic Motion Capture

Given a video:

- monocular (single view)
- without marker data
- uncalibrated camera
- in-the-wild

can we recover articulated **meshes** and **motion** in a **fixed** coordinate frame in 3D?



Why Monocular Motion Capture

- ✓ Industries directly interested:
  - Sports
  - Games/Animation
  - AR/VR
  - Autonomous driving
  - Fashion
- ✓ Behaviour modelling & understanding
- ✓ Solving such a highly-complex problem is interesting in itself (identifiability, occlusions, interlaced camera motion, in-the-wildness, etc)

Why Markerless: Less Effort & Money

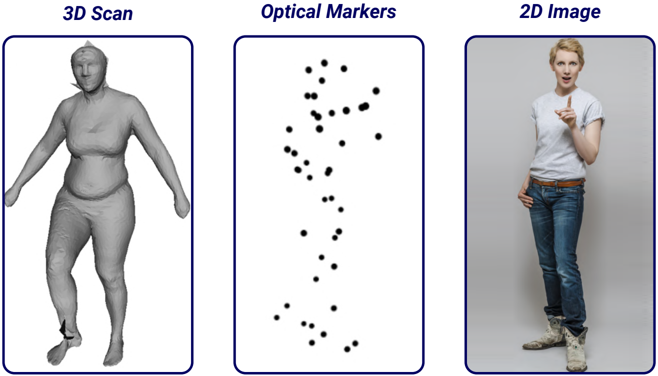


Figure 1: Possible input modalities for human motion capture, in decreasing complexity and equipment cost.

Why In The Wild: Increased Diversity

Figure 2: Randomly selected frames from scenes of 3DPW dataset[21].

Modelling Humans in 3D: Body Surface Only

[Ideally] model bones, joints, muscles, tissues, and skin (inside $\rightarrow$ out)

[Practice] only able to scan the outer body surface using 3D scanners

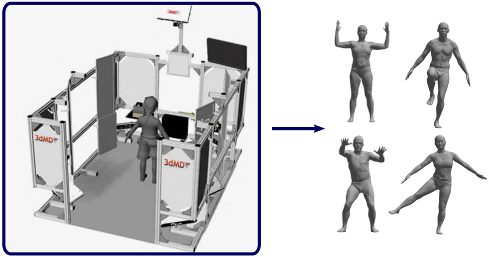
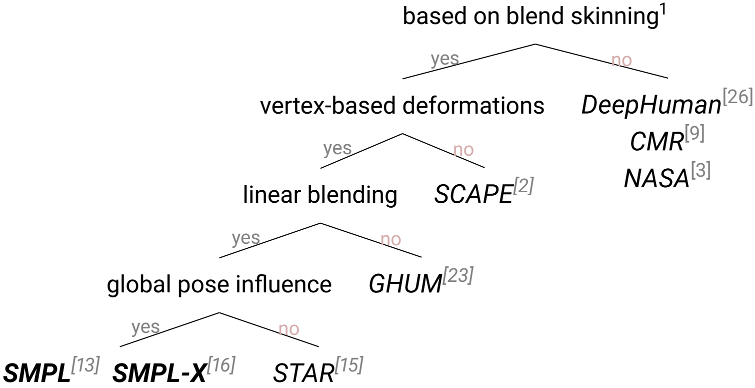


Figure 3: 3D body scanner from 3dMD (left). Registered meshes (right).

Modeling Humans in 3D: Overview of Methods



[1]Blend skinning is a skeleton-based deformation method, where mesh vertices are attached to joints via a set of weights[14].

Modeling Humans in 3D: The SMPL Parametric Model[13]

SMPL

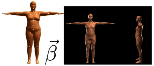
a learned model of human body shape and pose-dependent shape variation from 3D scans

- template-based
- blend-skinning
- linear in shape & pose-corrective[11] blend shapes
- differentiable

SMPL Model

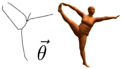
$$f(\vec{\boldsymbol{\theta}}, \vec{\boldsymbol{\beta}};\ \bar{\mathrm{T}}, S, P) \longmapsto \mathbb{R}^{3N_{\text{verts}}}$$

Shape: PCA coefficients



S = space of shape blend shapes

Pose: Rotation of 23 joints



P = space of pose corrective blend shapes

$\bar{\mathrm{T}}$ = average mesh

$\vec{\boldsymbol{\theta}}, \vec{\boldsymbol{\beta}}$

Modeling Humans in 3D: Fitting SMPL to Data

3D shape models enable the inference of object shape from noisy or ambiguous 2D/3D data.[5]

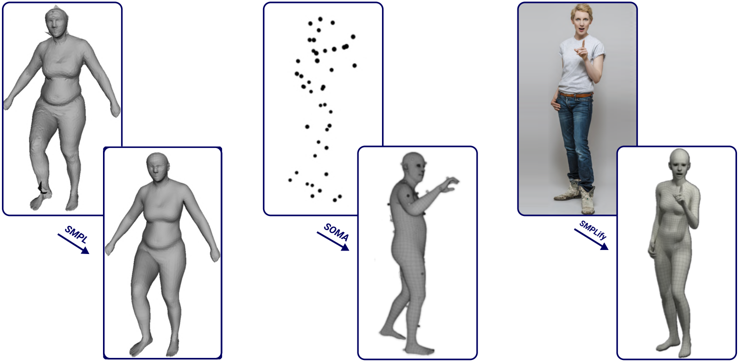


Figure 4: Fitting SMPL models to different input modalities.

Camera Is Moving: Motion Disentanglement Necessary

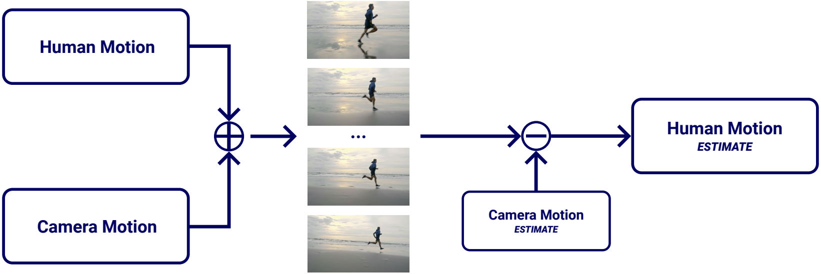
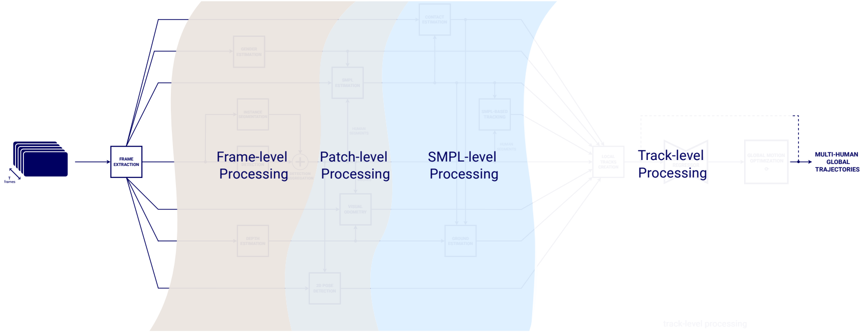
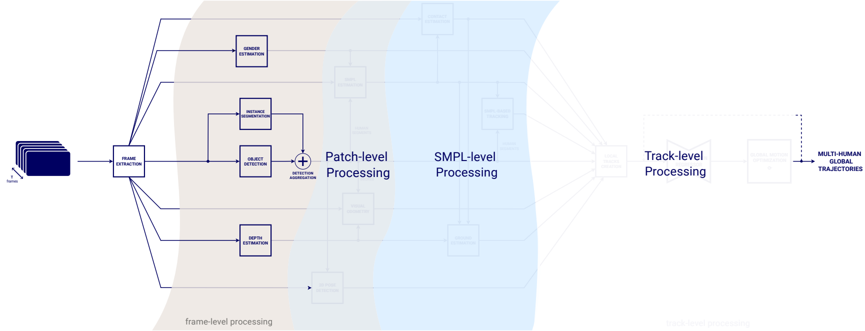


Figure 5: We observe the sum of human and camera motions. To disambiguate and recover human motion, camera's trajectory needs to be estimated.

Our 4-Stage System: Frame - Patch - SMPL - Track



Frame-Level Processing



Identifying Humans and Genders I

To identify humans in the input frames we employ the following visual learners:

- ***Co-DETR***[27] (236M params): based on Vision Transformers detects human subjects
- ***ConvNextV2***[22] (108M params): based on convnets detects and segments human instances
- ***MiVOLO***[10] (96M params): based on Vision Transformers detects human subjects and associated genders

The detections are *aggregated* based on IoU, resulting in robust human identification.

Identifying Humans and Genders II

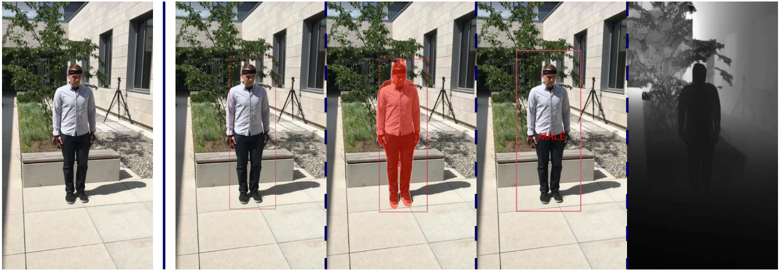
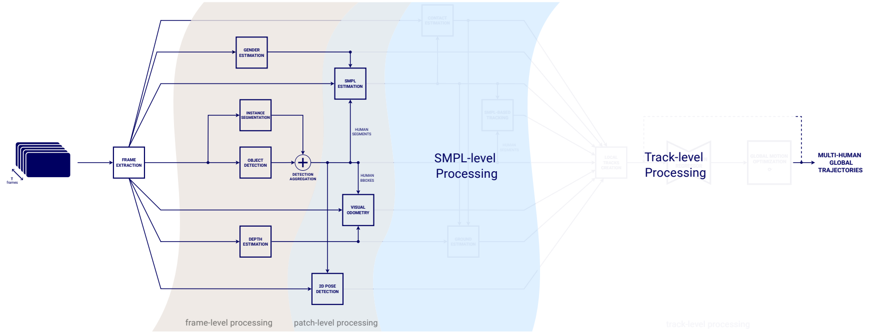


Figure 6: Output of frame-level processing nodes. From left to right: input, object detection, instance segmentation, gender, and depth estimation.

Patch-Level Operations



2D Pose Detection

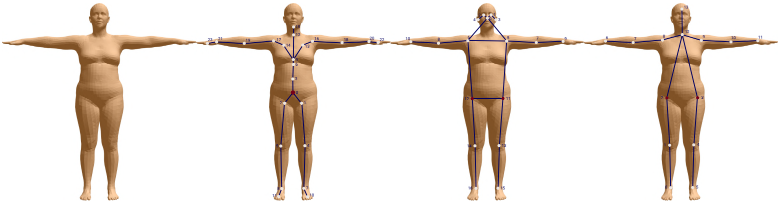


Figure 7: Different 2D pose formats: SMPL, COCO17[12], MPII14[1].

- Global motion recovery *sensitive* to detected 2D pose (main driving signal).
- ***ViTPose***[24] (308M params): based on Vision Transformers detects 17 keypoints (COCO17 format).

SMPL Parameters Estimation

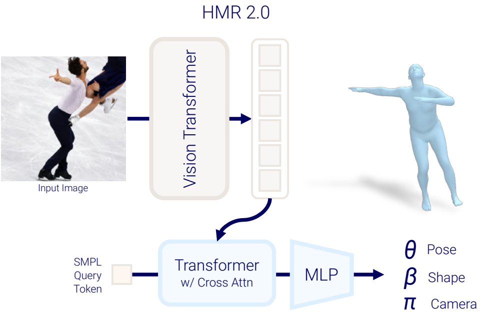


Figure 8: HMR 2.0[4], a network based on Vision Transformer for Human Mesh Recovery. θ, β are for SMPL, while π contains relative offsets from principal point and distance from camera.

Patch-Level Detections

Figure 9: Output of patch-level processing nodes. From left to right: female patch (squarified), estimated 2D pose COCO17 keypoints, projected SMPL mesh and joints. Subject isolation is performed for the male subject.

Human-Aware Visual Odometry

To estimate camera trajectory we use ***DPVO***[20] (7M params). We discourage sampling of patches inside human segmentation masks, while we pre-compute depth to ease matching process and BA.

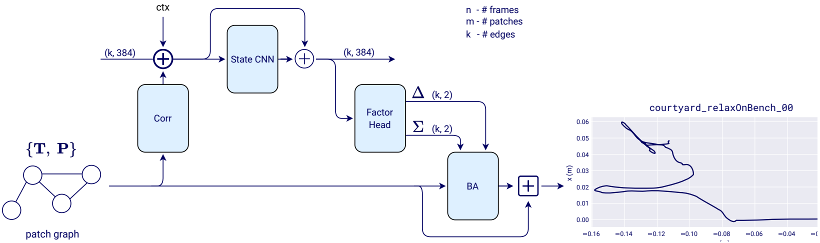
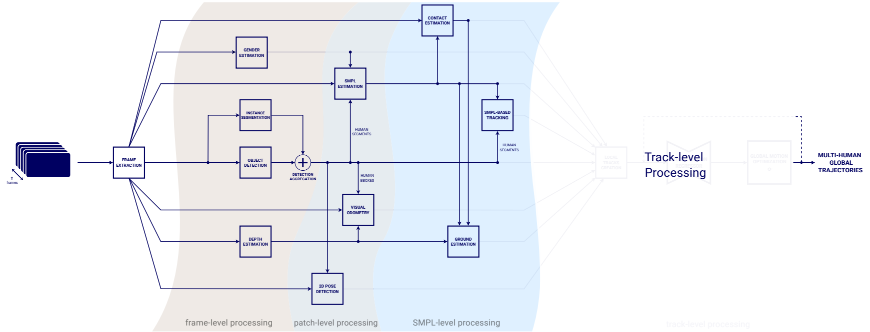


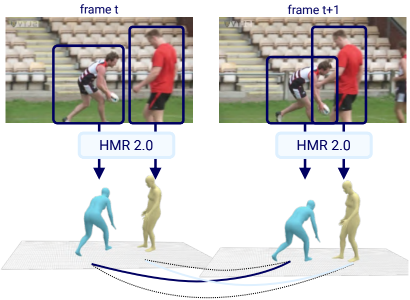
Figure 10: DPVO maintains a patch-to-frame association graph, which is processed to incrementally estimate camera poses. The output contains absolute values centered around the initial pose.

SMPL-Level Operations



Appearance-Aware Human Tracking

We use the ***PHALP***[17] (23M params) tracker to re-identify humans. Extends DeepSORT's state with SMPL pose, mesh location and regressed texture.



We achieve **1.08x** and **1.20x** IDs on 3DPW train and test sets.

Contacts and Ground Plane Estimation

To identify humans in the input frames we employ the following visual learners:

- ***BSTRO***[6] (243M params): based on SMPL pose and visual cues, it estimates per-vertex contact probabilities. We pool those using k-NN to get ankles and toes' contacts.
- **Ground Plane**: By clustering feet vertices in contact and iteratively merging planes we estimate local ground planes.

SMPL-Level Node Outputs

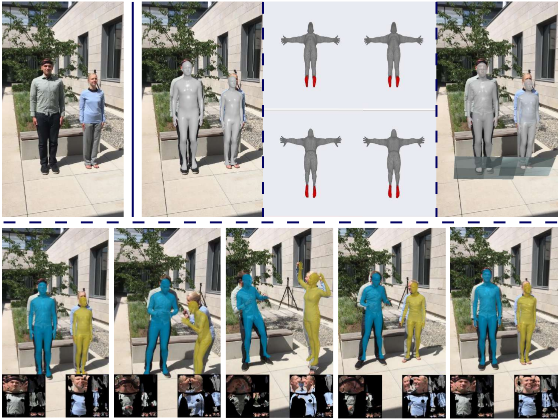


Figure 11: Top: input, SMPL fits, contacts, and ground estimation. Bottom: appearance-aware tracking and textures visualization.

1. Introduction
2. Pixel-Derived Information
3. **Human Motion Recovery**
4. Our Method
5. References

Camera-Local vs. Global Estimation

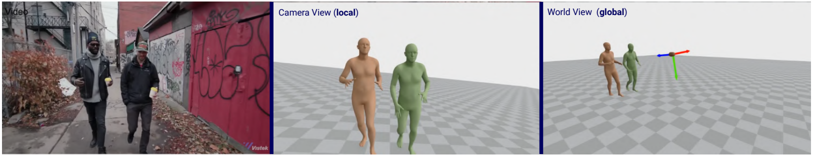


Figure 12: Input, camera-local, and global SMPL mesh placement [7].

The focus of this project, is *lifting* the camera-local human tracks to global frame.

Regression vs. Optimization Based Inference

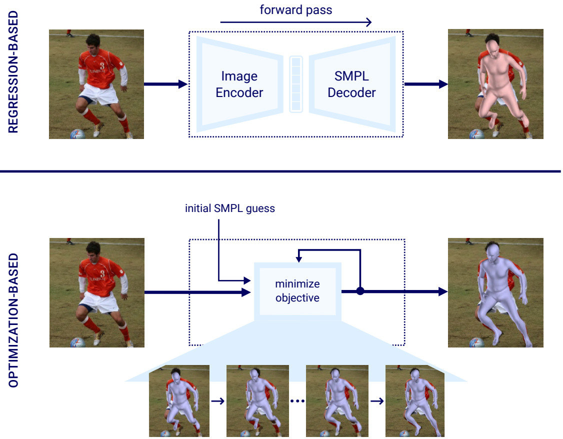


Figure 13: Estimating SMPL Parameters from monocular cues.

Combining Regression and Optimization for Local Inference

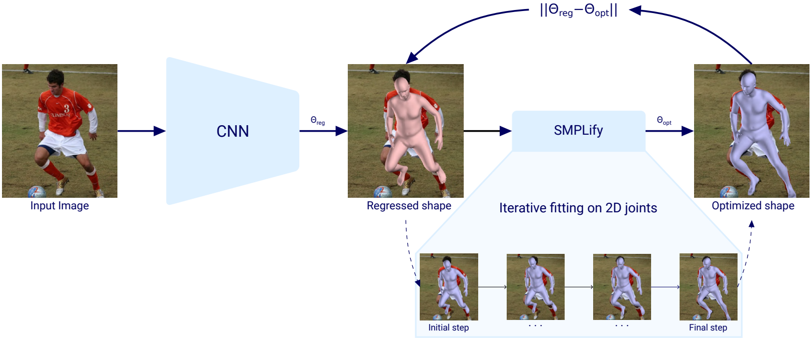


Figure 14: ***SPIN***[8] uses a visual regressor of SMPL parameters to initialize the hand-crafted optimization loop that follows.

Global Motion Recovery using Regression

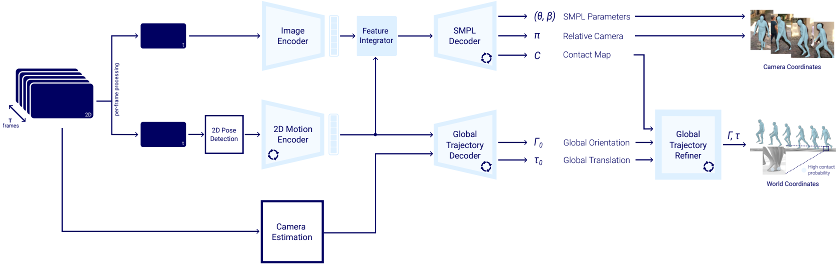


Figure 15: ***WHAM***[19] combines camera-local regression with motion disambiguation in an end-to-end trainable network, yielding SOTA motion results. Re-projection errors are high though.

In this work, we use a modified version of ***WHAM*** to initialize the global optimization loop.

Global Motion Recovery using Optimization

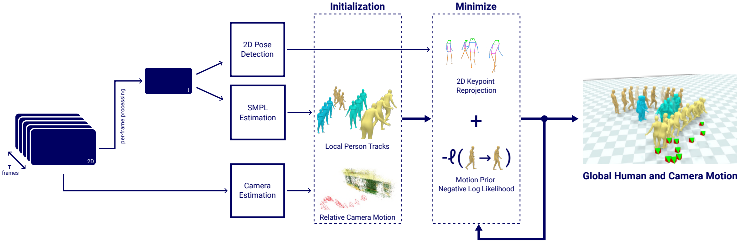


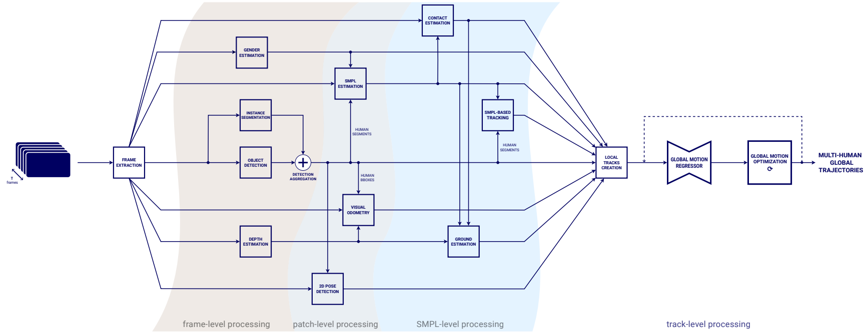
Figure 16: ***SLAHMR***[25] disentangles camera from human motion iteratively, by optimizing skeleton re-projection errors and motion realism[18]. Sensitive to initialization and loss formulation.

In this work, we use a global motion optimization recipe inspired from ***SLAHMR***.

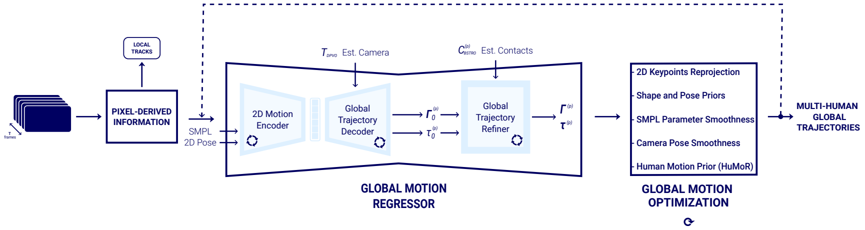
Methodology In A Nutshell

1. Identify Humans in the Video
   object detection, tracking, texture-based reID
2. Infer Camera-Local Human Meshes
   initial regression of ***SMPL*** parameters
3. Infer Camera Poses
   single camera (monocular), visual odometry
4. Lift Local Tracks to Global
   motion semantics, re-projection consistency, global motion smoothness

Methodology In Detail



Global Motion Lifting Using Regression and Optimization



3D Poses In The Wild (3DPW)

- Challenging capturing setups with moving cameras
- Diverse human actions (e.g. walking, running, climbing, arguing)
- 61 scenes, 51K frames with ground-truth IMU and SMPL annotations

Figure 17: Sample frame from 3DPW (left) and visualization of 2D poses alongside camera-accurate, textured SMPL meshes.

Evaluation Metrics I

1. **Mean Per-Joint Position Error (MPJPE)**: average euclidean distance (mm) between predicted and ground-truth joint positions.

$$E_{MPJPE} = \frac{1}{N_J} \sum_{i=1}^{N_J} J_i^{gt} - J_i^{pred\,2}$$

We compute this metric using $N_J = 14$ joints following MPII[1].

2. **Procrustes Aligned Mean Per-Joint Position Error (PA-MPJPE)**: accounts for potential rigid transformations between the prediction and ground truth by estimating a similarity transformation before calculating the error.

Evaluation Metrics II

3. **Mean Per-Vertex Error (PVE)**: average position error (mm) of the mesh vertices

$$E_{PVE} = \frac{1}{N_V} \sum_{i=1}^{N_V} V_i^{gt} - V_i^{pred\,2}$$

4. **Acceleration Error (Acc)**: average difference (m/s²) of the joint accelerations (computed using the recording FPS)

$$A_j^{gt} = J_{j,1:T_p-2}^{gt} - 2 \times J_{j,2:T_p-1}^{gt} + J_{j,3:T_p}^{gt}$$

$$A_j^{pred} = J_{j,1:T_p-2}^{pred} - 2 \times J_{j,2:T_p-1}^{pred} + J_{j,3:T_p}^{pred}$$

$$Acc = \frac{1}{N_J} \sum_{i=1}^{N_J} A_i^{gt} - A_i^{pred\,2} \times fps^2$$

Evaluation on 3DPW

Models	3DPW			
	PA-MPJPE	MPJPE	PVE	Acc
HMR2.0 [4] *	44.4	69.8	82.2	18.1
WHAM [19] *	37.2	59.4	71.0	6.9
SLAHMR [25] *	55.9	-	-	-
Local	48.1	70.2	90.8	17.8
Global Regr	**43.1**	74.5	101.1	**7.2**
Global Opt	49.7	**59.9**	**74.4**	9.1
Global Regr + Opt	43.4	64.1	74.3	7.6

Table 1: Global motion estimation metrics on 3DPW [21] aggregated across all dataset scenes. The metrics denoted with * are taken from original papers.

Qualitative Results

Figure 18: Visualization of the recovered human motion. The backgrounds corresponding to the last frame are used; transparency indicates track age.

Discussion

- Qualitative results highlight the necessity of lifting local tracks to a fixed global frame.
- Global lifting using regression (b) results in smooth motions sacrificing consistency with visual cues.
- Further optimization of the global trajectory (c) results in better projection consistency and smoother mesh translations relative to the world frame.
- **Regression** enables efficient initial global motion estimate, while hand-crafted **optimization** reduces non-PA reconstruction errors by optimizing re-projection and motion criteria. We have showed that their **combination leads to effective global motion recovery**.

[1] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele.
2d human pose estimation: New benchmark and state of the art analysis.
In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693, 2014.

[2] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis.
Scape: shape completion and animation of people.
In *ACM SIGGRAPH 2005 Papers*, pages 408–416. 2005.

[3] B. Deng, J. P. Lewis, T. Jeruzalski, G. Pons-Moll, G. Hinton, M. Norouzi, and A. Tagliasacchi.
Nasa neural articulated shape approximation.
In *European Conference on Computer Vision*, pages 612–628. Springer, 2020.

[4] S. Goel, G. Pavlakos, J. Rajasegaran, A. Kanazawa, and J. Malik.
Humans in 4d: Reconstructing and tracking humans with transformers.
arXiv preprint arXiv:2305.20091, 2023.

[5] D. Hirshberg, M. Loper, E. Rachlin, and M. Black.
Coregistration: Simultaneous alignment and modeling of articulated 3D shape.
In *European Conf. on Computer Vision (ECCV)*, LNCS 7577, Part IV, pages 242–255. Springer-Verlag, Oct. 2012.

[6] C.-H. P. Huang, H. Yi, M. Höschle, M. Safroshkin, T. Alexiadis, S. Polikovsky, D. Scharstein, and M. J. Black.
Capturing and inferring dense full-body human-scene contact.
In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 13274–13285, June 2022.

[7] M. Kocabas, Y. Yuan, P. Molchanov, Y. Guo, M. J. Black, O. Hilliges, J. Kautz, and U. Iqbal.
Pace: Human and camera motion estimation from in-the-wild videos.
arXiv preprint arXiv:2310.13768, 2023.

[8] N. Kolotouros, G. Pavlakos, M. J. Black, and K. Daniilidis.
Learning to reconstruct 3d human pose and shape via model-fitting in the loop.
In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2252–2261, 2019.

[9] N. Kolotouros, G. Pavlakos, and K. Daniilidis.
Convolutional mesh regression for single-image human shape reconstruction.
In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4501–4510, 2019.

[10] M. Kuprashevich and I. Tolstykh.
Mivolo: Multi-input transformer for age and gender estimation.
arXiv preprint arXiv:2307.04616, 2023.

[11] J. P. Lewis, M. Cordner, and N. Fong.
Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation.
In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 165–172, 2000.

[12] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick.
Microsoft coco: Common objects in context.
In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[13] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black.
Smpl: A skinned multi-person linear model.
ACM Trans. Graph., 34(6), nov 2015.

[14] N. Magnenat-Thalmann, R. Laperrire, and D. Thalmann.
Joint-dependent local deformations for hand animation and object grasping.
In *In Proceedings on Graphics interface'88*. Citeseer, 1988.

[15] A. A. A. Osman, T. Bolkart, and M. J. Black.
STAR: A sparse trained articulated human body regressor.
In *European Conference on Computer Vision (ECCV)*, pages 598–613, 2020.

[16] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas, and M. J. Black.
Expressive body capture: 3D hands, face, and body from a single image.
In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019.

[17] J. Rajasegaran, G. Pavlakos, A. Kanazawa, and J. Malik.
Tracking people by predicting 3d appearance, location and pose.
In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2740–2749, 2022.

[18] D. Rempe, T. Birdal, A. Hertzmann, J. Yang, S. Sridhar, and L. J. Guibas.
Humor: 3d human motion model for robust pose estimation.
page 11468–11479. IEEE Computer Society, 2021.

[19] S. Shin, J. Kim, E. Halilaj, and M. J. Black.
Wham: Reconstructing world-grounded humans with accurate 3d motion.
arXiv preprint arXiv:2312.07531, 2023.

[20] Z. Teed, L. Lipson, and J. Deng.
Deep patch visual odometry.
Advances in Neural Information Processing Systems, 36, 2024.

[21] T. Von Marcard, R. Henschel, M. J. Black, B. Rosenhahn, and G. Pons-Moll.
Recovering accurate 3d human pose in the wild using imus and a moving camera.
In *Proceedings of the European conference on computer vision (ECCV)*, pages 601–617, 2018.

[22] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie.
Convnext v2: Co-designing and scaling convnets with masked autoencoders.
In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16133–16142, 2023.

[23] H. Xu, E. G. Bazavan, A. Zanfir, W. T. Freeman, R. Sukthankar, and C. Sminchisescu.
Ghum & ghuml: Generative 3d human shape and articulated pose models.
In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6184–6193, 2020.

[24] Y. Xu, J. Zhang, Q. Zhang, and D. Tao.
Vitpose: Simple vision transformer baselines for human pose estimation.
Advances in Neural Information Processing Systems, 35:38571–38584, 2022.

[25] V. Ye, G. Pavlakos, J. Malik, and A. Kanazawa.
Decoupling human and camera motion from videos in the wild.
In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21222–21232, 2023.

[26] Z. Zheng, T. Yu, Y. Wei, Q. Dai, and Y. Liu.
Deephuman: 3d human reconstruction from a single image.
In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7739–7749, 2019.

[27] Z. Zong, G. Song, and Y. Liu.
Detrs with collaborative hybrid assignments training.
In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6748–6758, 2023.

Thank you for attending!

ACKNOWLEDGMENTS

Prof. Hedvig Kjellström for her guidance, support, and fruitful discussions.

Prof. Jonas Beskow for reviewing and examining this work.

My **parents**, **siblings**, and **friends** for supporting me throughout my academic journey.