Two Views Are Better than One: Monocular 3D Pose Estimation with Multiview Consistency

Christian Keilstrup Ingwersen[1], Rasmus Tirsgaard[2], Rasmus Nylander[1], Janus Nørtoft Jensen[3], Anders Bjorholm Dahl[2], and Morten Rieger Hannemose[2]

Trackman A/S[1]

Technical University of Denmark[2]

Presented by

DaeYong Kim, Dept. of Artificial Intelligence, Ajou University

Introduction & Background

Objective

- The challenge of estimating 3D human pose from a single 2D image

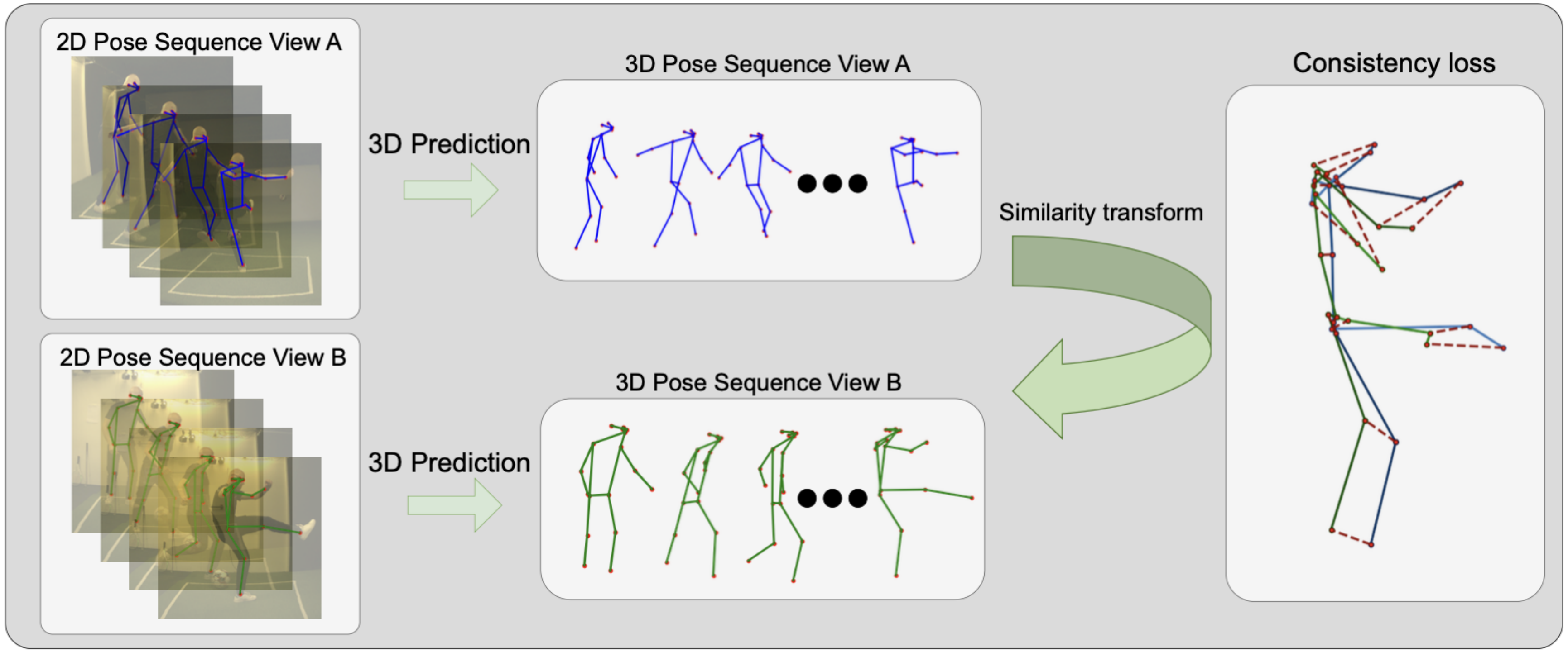
Limitations of Previous Methods

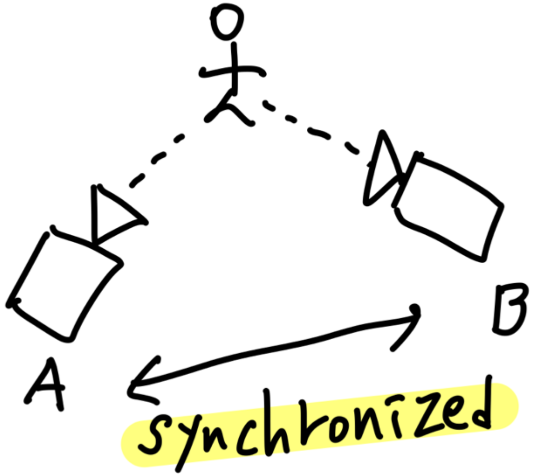
- 3D data is accurate but expensive and hard to obtain
- 2D data is abundant but lacks of depth information, depth ambiguity issue

Method

Consistency Loss

- Enforces consistency between 3D poses inferred from different views
- Uses Procrustes analysis to align 3D posed from Multiple views
- Without camera calibration, extrinsic/intrinsic parameters not needed

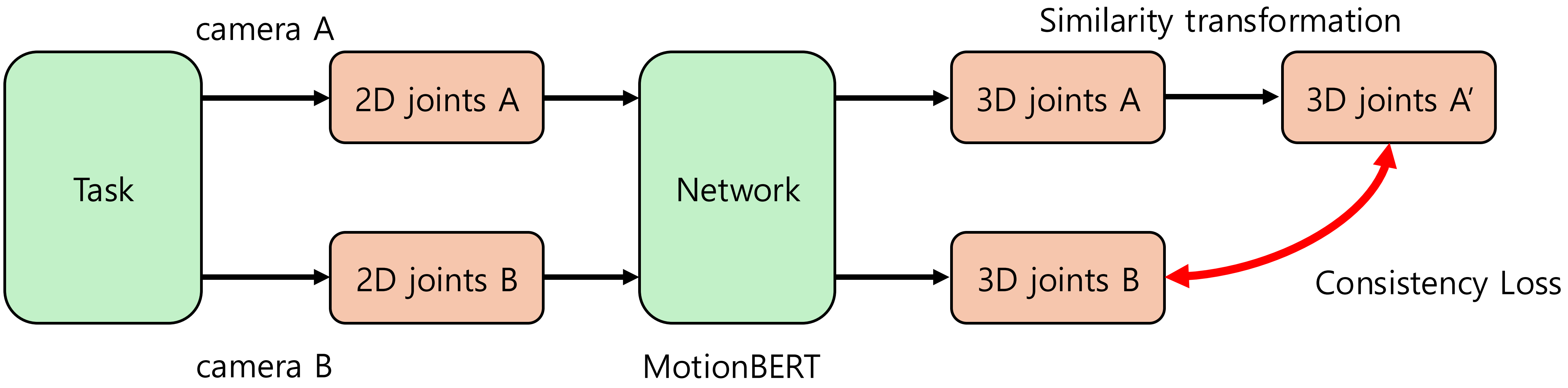




<Method Overview>

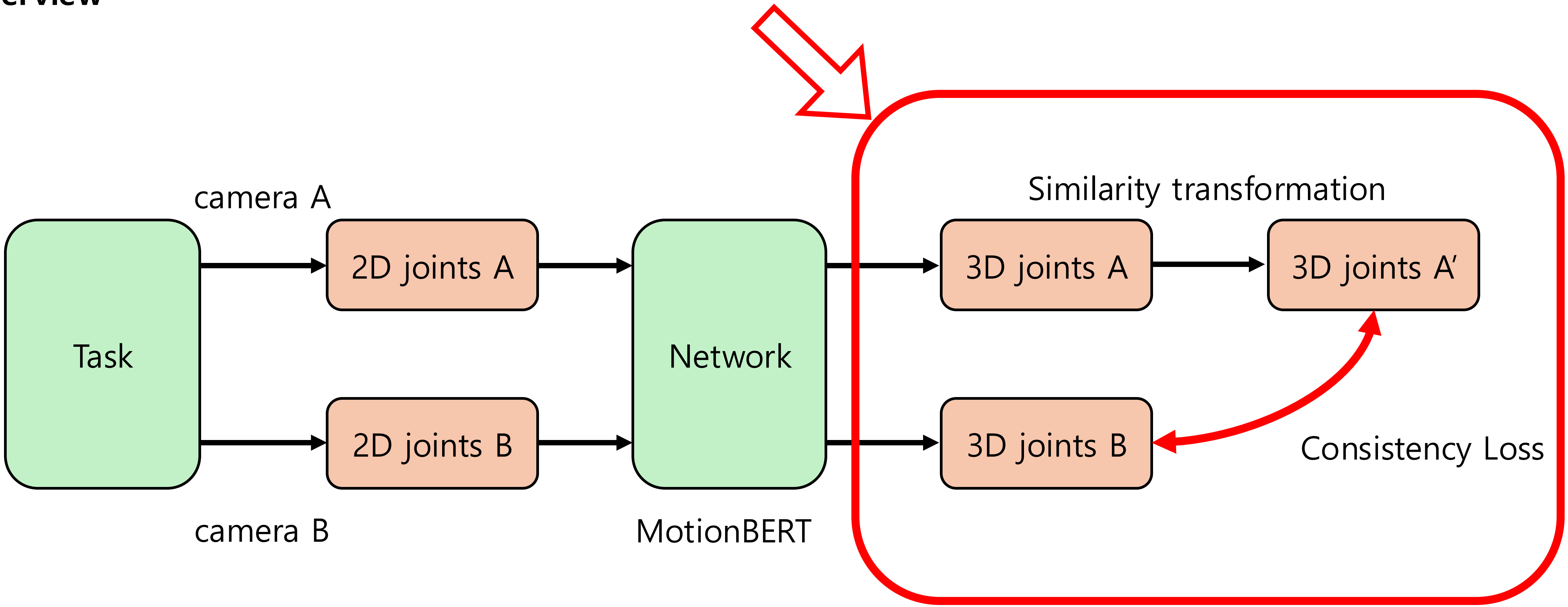
Method

Overview



Method

Overview



Method

Similarity transformation

- It is a geometric transformation that reserves the shape of an object

Scaling, Rotation, Translation

$$X' = sRX + t$$

Why is Similarity transformation

- 3D poses predicted from different camera views are in different coordinate systems.
- Directly comparing them is difficult because of **scale, rotation, and position differences.**
- To eliminate the need for camera calibration (extrinsic/intrinsic parameters).

Method

Similarity transformation

- It is a geometric transformation that reserves the shape of an object

Scaling, Rotation, Translation

- we have to calculate the optimal similarity transform with parameters $\hat{\theta}_{ab}$

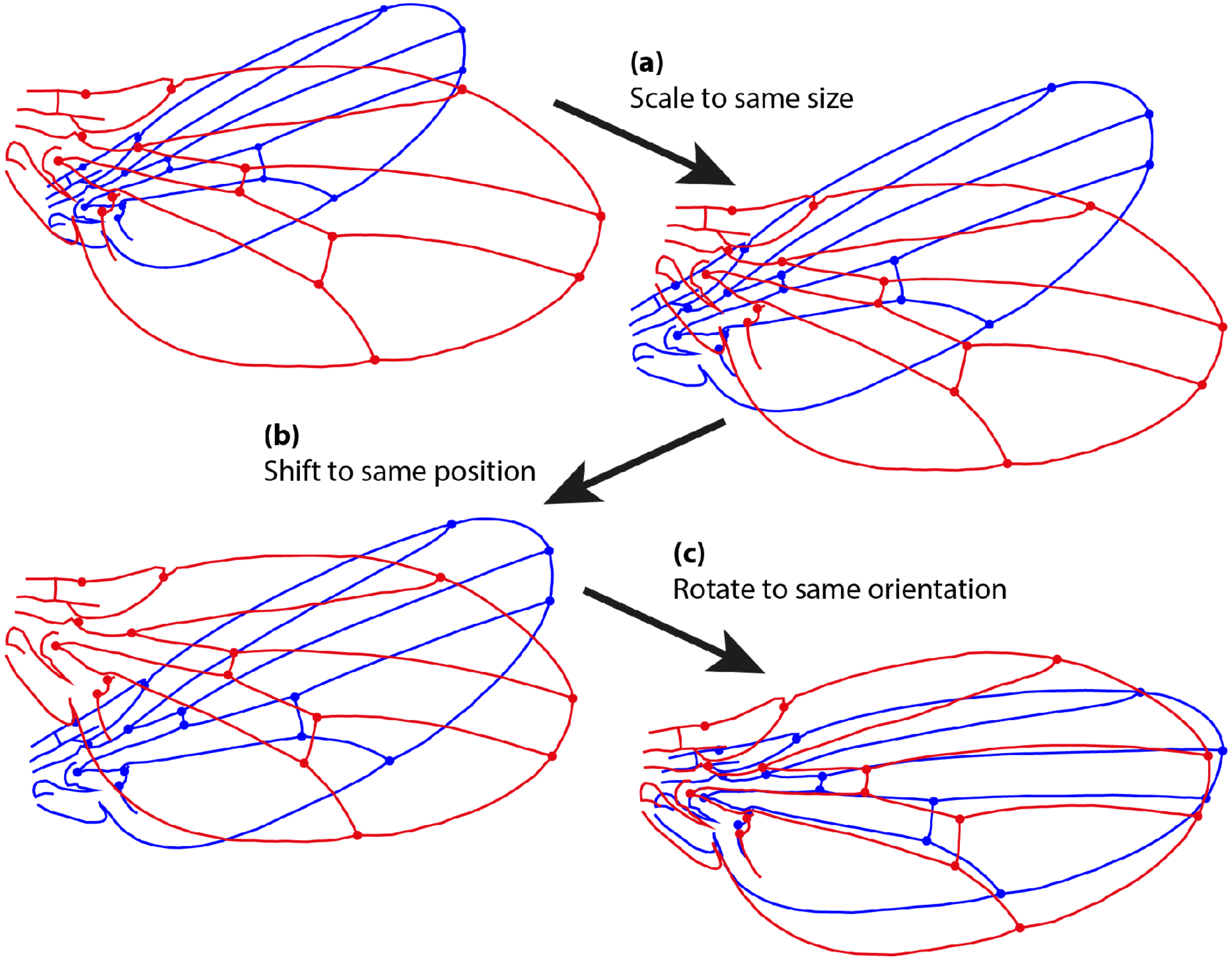
$$\hat{\theta}_{ab} = \arg\min_{\theta} \sum_{i=1}^{n} \left|\left|\tau\left(\hat{J}_{a,i};\theta\right) - \hat{J}_{b,i}\right|\right|_2^2$$

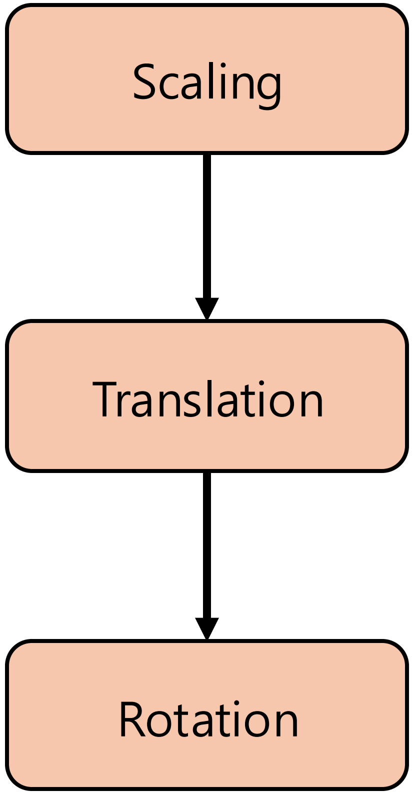
$$\tau\left(\hat{J}_{a,i};\hat{\theta}_{ab}\right) = s\hat{J}_{a,i}R + t$$

Method

Procrustes Analysis

- It it a form of statistical shape analysis used to analyze the distribution of a set of shapes.





Method

mean of Consistency Loss

- The mean difference over every pair of two cameras

S : the total of sequences,

V: the set of possible pairs of views of the sequence

$$\mathcal{L}_{\text{con}} = \sum_{s=1}^{S} \frac{1}{|V_s|} \sum_{(a,b)\in V_s} \mathcal{L}_{\text{c}}\left(\hat{J}_a, \hat{J}_b\right)$$

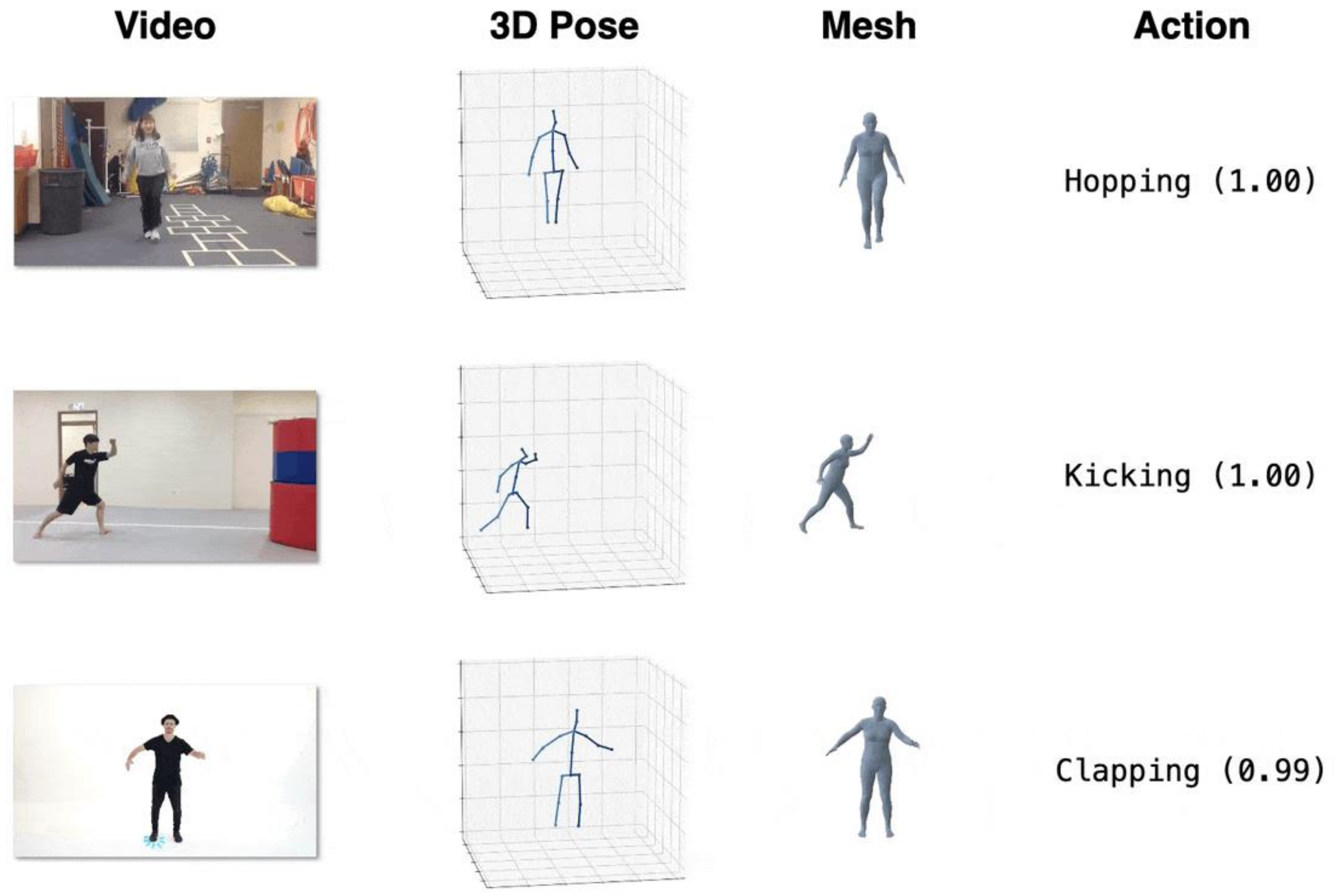
Consistency Loss

$$\mathcal{L}_{\text{c}}(\hat{J}_a, \hat{J}_b) = \frac{1}{n} \sum_{i=1}^{n} \left\| \tau\left(\hat{J}_{a,i}; \hat{\theta}_{ab}\right) - \hat{J}_{b,i} \right\|_2$$

Experimental Setup & Datasets

Used Model

- motionBERT: 2D → 3D Pose

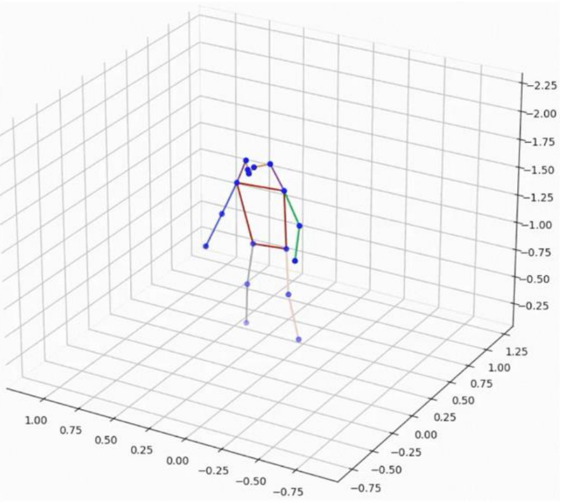


MotionBERT: A Unified Perspective on Learning Human Motion Representations (ICCV 2023)

Experimental Setup & Datasets

Datasets

- SportsPose: Dynamic sports movements, additional views included (fine-tuning)
- Human3.6M: Used for semi-supervised learning experiments (semi-supervised learning)



https://christianingwersen.github.io/SportsPose/
http://vision.imar.ro/human3.6m/description.php

Evaluations

Fine-tuning: SportsPose

	Soccer kick		Tennis serve		Baseball pitch		Volley		Jumping		All	
	MPJPE	PA	MPJPE	PA	MPJPE	PA	MPJPE	PA	MPJPE	PA	MPJPE	PA
Baseline												
MotionBERT [46]	64.2	39.5	70.7	39.7	85.0	42.2	86.8	50.0	78.0	48.9	77.1	44.1
Iqbal *et al.* [14][4]	42.8	28.5	39.9	26.2	47.0	30.1	40.3	27.9	44.9	30.1	42.9	28.5
Fine-tuning with 3D data (2 views)												
$\mathcal{L}_{3D}$ (5)	26.7	**20.2**	27.3	20.1	30.1	22.5	31.3	24.2	27.9	21.4	28.7	21.7
$\mathcal{L}_{3D_{con}}$ (6), Ours	**26.1**	20.5	**25.4**	**18.8**	**29.4**	**22.4**	**30.8**	**23.8**	**27.9**	**20.9**	**28.0**	**21.3**
Only 2D fine-tuning (2 views)												
$\mathcal{L}_{2D}$ (7)	59.0	44.1	59.1	42.0	73.8	45.1	64.7	47.8	65.0	45.6	64.4	45.0
$\mathcal{L}_{2D_{con}}$ (8), Ours	**36.6**	**22.5**	**34.2**	**22.2**	**41.7**	**25.2**	**37.3**	**23.7**	**32.0**	**22.7**	**36.4**	**22.0**

<Evaluation Table>

Right + view x	MPJPE	PA-MPJPE
View 1	21.8	**22.4**
View 2	**21.6**	24.3
View 3	27.3	31.8
View 4	25.6	26.7
View 5	31.9	35.6
View 6	25.8	27.2

<Which views to use>

Evaluations

Semi-supervised: Human3.6M

- 3D data: supervised learning
- 2D data: Consistency Loss (No-labels)

Methods	MPJPE↓	PA-MPJPE↓
Rodhin et al. (ECCV'18) [35]	131.7	98.2
Pavlakos et al. (ICCV'19) [32]	110.7	74.5
Li et al. (ICCV'19) [22]	88.8	66.5
Rodhin et al. (CVPR'18) [36]	-	65.1
Kocabas et al. (CVPR'19) [20]	-	60.2
Iqbal et al. (CVPR'20) [14]	62.8	51.4
Roy et al. (3DV'22) [37]	60.8	48.4
Ours	**58.9**	**43.6**

<Evaluation Table>

Limitations & Contributions

Limitations

- Performance depends on camera placements
- Requires fixed camera positions
- Requires precise camera synchronization

Contributions

- Works without camera calibration.
- Significantly improves performance even without 3D ground truth data.

Q&A