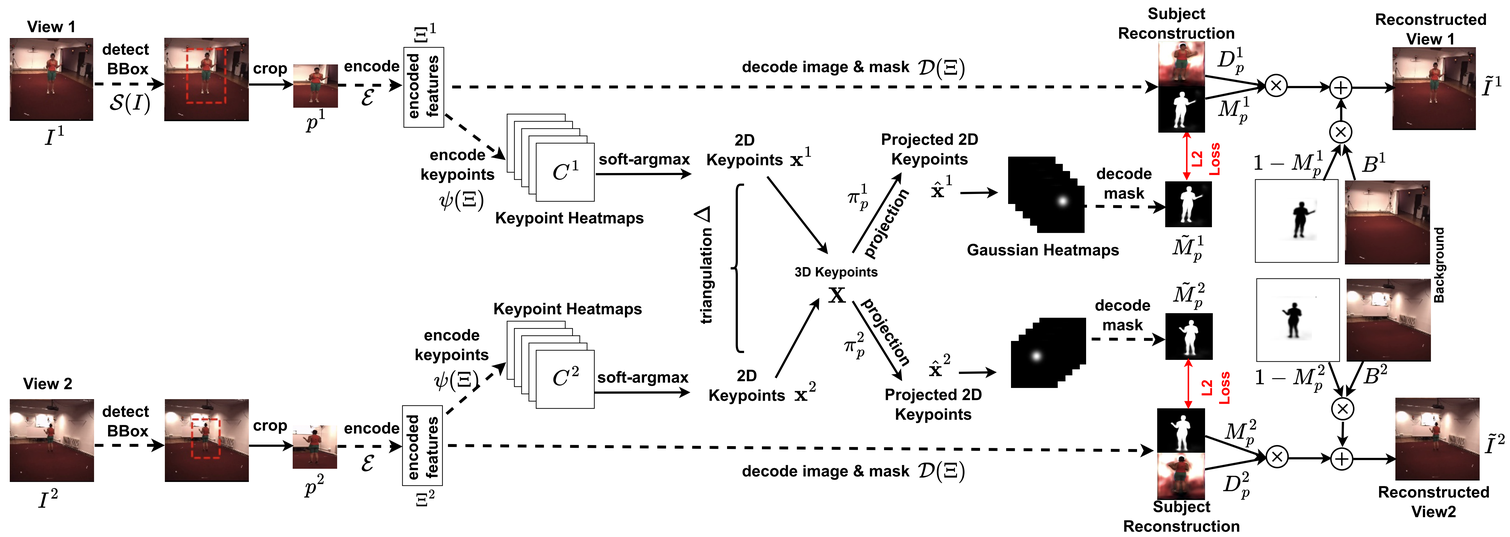
Unsupervised 3D Keypoint Discovery with Multi-View Geometry

Sina Honari, Chen Zhao, Mathieu Salzmann, Pascal Fua

EPFL CVLab

Motivation:

3D human pose estimation models require a considerable amount of labels, which is difficult to obtain. In this paper, we propose a method that discovers unsupervised 3D keypoints. These keypoints can be later mapped to the target 3D pose of interest with a small set of labels (using a simple MLP).



Method:

- Subject is first detected and cropped and together with the background image reconstruct the input image through subject mask estimation.

$$\tilde{\mathbf{I}} = \mathbf{M} \times \mathbf{D} + (1 - \mathbf{M}) \times \mathbf{B}$$

- Encoded features $\Xi^{\mathbf{v}}$ from each view v are are used to predict 2D keypoints $\mathbf{x}$ in each view, which are then projected to 3D using a triangulation operation Δ

$$\Delta(\{\mathbf{x}_n^{\vartheta}\}_{\vartheta=1}^{V}, \{\mathbf{\Pi}_{\mathbf{p}}^{\vartheta}\}_{\vartheta=1}^{V}) = \mathbf{X}_n$$

- The 3D keypoints $\mathbf{X}$ are first projected to 2D in each view and then used to estimated the subject mask $\mathbf{M}$ that the model itself has initially estimated.

$$\tilde{\mathbf{M}}_{\mathbf{p}}^{\vartheta} = \phi(\{\hat{\mathbf{x}}_n^{\vartheta}\}_{n=1}^{N})$$

- **No label or pretrained model is used for subject detection, mask estimation, 2D, or 3D keypoint estimation.**

Losses:

- **Image reconstruction loss**:

$$\mathcal{L}_{\text{reconst}} = \|\mathbf{I} - \tilde{\mathbf{I}}\|_2^2 + \beta \sum_{l=1}^{3} \|\text{Res}_l(\mathbf{I}) - \text{Res}_l(\tilde{\mathbf{I}})\|_2^2$$

- **Mask reconstruction loss**: mask of the keypints path should match the mask of the image reconstruction path

$$\mathcal{L}_{\text{mask}} = \|\tilde{\mathbf{M}}_{\mathbf{p}} - \mathbf{M}_{\mathbf{p}}\|_2^2$$

- **Coverage loss**: the keypoints should fall on the subject mask

$$\mathcal{L}_{\text{coverage}} = \frac{1}{n} \sum_{n=1}^{N} |1 - \bar{\mathbf{H}}_n \odot \mathbf{M}_{\mathbf{p}}|$$

- **Centering loss**: the center of bounding box (bbox) should be close to the average of keypoints (robust bbox pred)

$$\mathcal{L}_{\text{centering}} = |\mathbf{u} - \frac{1}{N} \sum_{n=1}^{N} \hat{\mathbf{x}}_n|$$

- **All losses**:

$$\mathcal{L}_{\text{unsup}} = \mathcal{L}_{\text{reconst}} + \gamma \mathcal{L}_{\text{mask}} + \delta \mathcal{L}_{\text{coverage}} + \eta \mathcal{L}_{\text{centering}}$$

Results:

Table 1. Comparison with unsupervised 3D models on H36M (in mm). SV and MV indicate single- and multi-view models. kpts denotes number of keypoints.

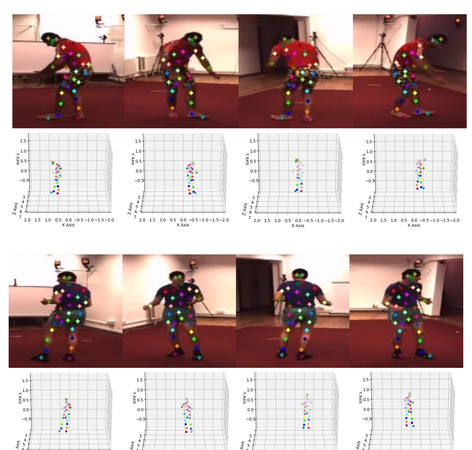
Model				MPJPE	N-MPJPE	P-MPJPE
Known Kinematic Model						
Kundu *et al.*[26]				99.2	-	-
Kundu *et al.*[27]				-	-	89.4
Uninterpretable latents						
NVS [35]				-	115.0	-
Honari *et al.*[15]				100.3	99.3	74.9
Keypoint Discovery						
KeypointNet [41]	SV	2 hid MLP	32 kpts	158.7	156.8	112.9
Ours	SV	2 hid MLP	32 kpts	**125.73**	**121.04**	**89.05**
BKinD-3D [40]	MV	Linear	15 kpts[a]	125	-	105
Ours	MV	Linear	32 kpts	**120.9**	**117.9**	**93.5**
Ours	MV	2 hid MLP	32 kpts	**73.8**	**72.6**	**63.0**

Table 2. Comparison with unsupervised 3D models on 3DHP (in cm). All models use a 2 hidden layer MLP to map either features (top 3 models) or 48 discovered keypoints (Ours) to the labelled pose.

Model	Train-Set	MPJPE	N-MPJPE	P-MPJPE
Uninterpretable latents				
DrNet [7]	3DHP	22.28	21.55	14.94
NSD [37]	3DHP	20.24	19.29	14.09
Honari *et al.* [15]	3DHP	20.95	19.78	14.04
Keypoint Discovery				
Ours	H36M	16.90	16.19	12.48
Ours	3DHP	**14.57**	**14.21**	**11.52**

Table 3. Comparison with 2D keypoint estimation models. All models predict 32 keypoints on 6 actions of wait, pose, greet, direct, discuss, and walk and regress a linear model from 2D keypoints to the 2D pose labels.

Model	%-MSE Error
Thewlis et al. [43]	7.51
Zhang et al. [49]	4.14
Schmidtke et al. [38]	3.31
Lorenz et al. [30]	2.79
Jakab et al. [21]	2.73
Ours	**2.38**



Ablation studies:

$\mathcal{L}_{\text{reconst}}$	$\mathcal{L}_{\text{mask}}$	$\mathcal{L}_{\text{coverage}}$	$\mathcal{L}_{\text{centering}}$	MPJPE	N-MPJPE	P-MPJPE
✓	✗	✗	✗	111.8	107.6	79.7
✓	✓	✗	✗	79.9	78.6	67.0
✓	✓	✓	✗	78.3	77.0	65.6
✓	✓	✓	✓	**73.8**	**72.6**	**63.0**

Number of Views	Pose Model	MPJPE	N-MPJPE	P-MPJPE
2	2-hid MLP	103.21	100.7	81.6
3	2-hid MLP	77.7	75.7	64.0
4	2-hid MLP	**73.8**	**72.6**	**63.0**