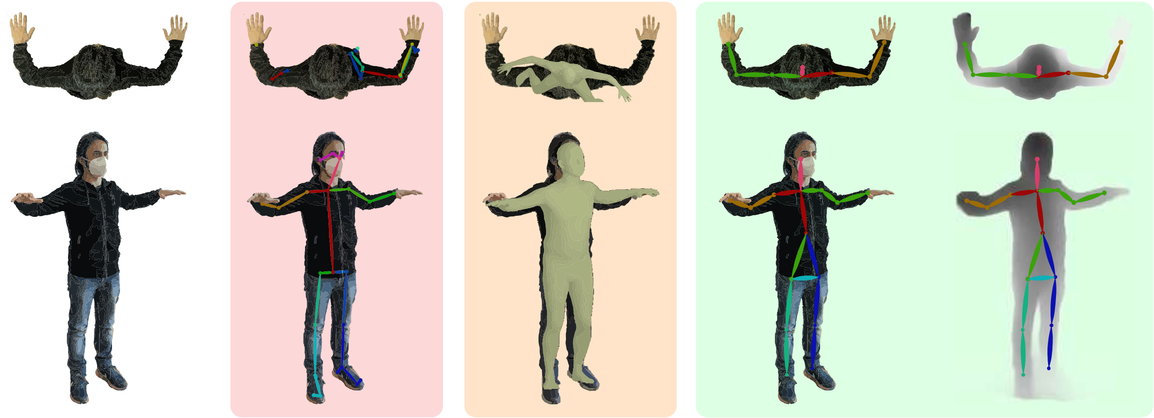
UNIVERSITY
OF TRENTO - Italy

DECA: Deep viewpoint-Equivariant human pose estimation using Capsule Autoencoders

Nicola Garau[1] Niccolò Bisagno[1] Piotr Bródka[1] Nicola Conci[1]

[1]University of Trento, Via Sommarive, 9, 38123 Povo, Trento TN
{nicola.garau, niccolo.bisagno}@unitn.it, piotrbrodka95@gmail.com, nicola.conci@unitn.it

Overview



Input (a) 2D/3D HPE (b) 3D HMR (c) Ours

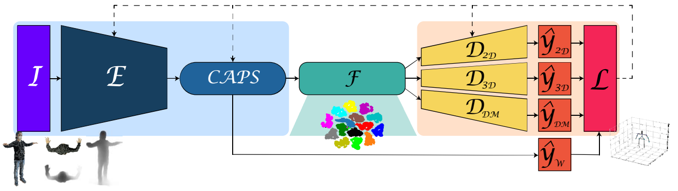
Context

Current 3D HPE methods suffer a lack of viewpoint equivariance, namely they tend to fail or perform poorly when dealing with viewpoints unseen at training time.

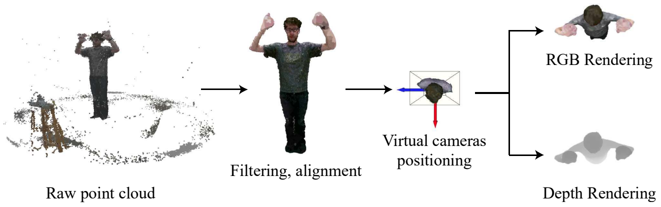
We propose DECA, a novel capsule autoencoder network that allows to drastically reduce the network data dependency at training time, resulting in an improved ability to generalize to unseen viewpoints.

$$\mathcal{L} = \sum_{\tau \in \mathcal{T}} \left(s_\tau + e^{-s_\tau} \mathcal{L}_\tau \right) \qquad \mathcal{L}_\tau = \begin{cases} \mathcal{L}_{2\mathcal{D},3\mathcal{D}} = \frac{1}{BS} \sum_{i=0}^{BS} (y_i - \hat{y}_i)^2 \\ \mathcal{L}_{\mathcal{DM}} = \dfrac{\sum_{i=0}^{BS} \left[mask\,|y_i - \hat{y}_i| + |y_i - \hat{y}_i| \right]}{2 * BS} \\ \mathcal{L}_{\mathcal{W}} = \|\hat{y}_{\mathcal{W}} W_{ij}\|_F \end{cases}$$

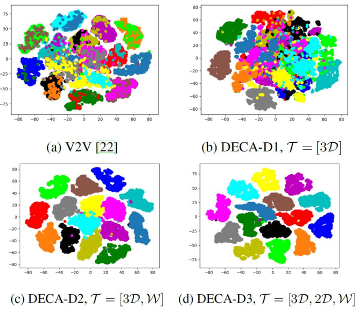
Network



PanopTOP dataset [2]



Latent space and results

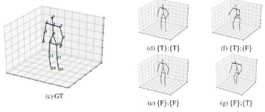


(a) V2V [22] (b) DECA-D1, $\mathcal{T} = [3\mathcal{D}]$

(c) DECA-D2, $\mathcal{T} = [3\mathcal{D}, \mathcal{W}]$ (d) DECA-D3, $\mathcal{T} = [3\mathcal{D}, 2\mathcal{D}, \mathcal{W}]$

	ITOP Train on front, test on top				
Body part	RF [28]	RTW [37]	IEF [3]	VI [9]	DECA-D3
Head	48.10	1.50	47.90	55.60	46.27
Neck	5.90	8.10	39.00	40.90	**73.14**
Torso	4.70	3.90	41.90	35.00	**85.94**
Upper Body	19.70	2.20	23.90	29.40	**45.00**
Full Body	10.80	2.00	17.40	20.40	**51.85**

Table 2: Comparison with the state-of the art for the ITOP viewpoint transfer task (metric: 0.1m mAP). Training on front-view, validating on front-view, testing on top-view (top-view data is unseen in validation).



(c) GT (d) {T};{T} (f) {T};{F} (e) {F};{F} (g) {F};{T}

[1] Garau, N., Bisagno, N., Bródka, P., & Conci, N. (2021). DECA: Deep viewpoint-Equivariant human pose estimation using Capsule Autoencoders. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 11677-11686).

[2] Garau, N., Martinelli, G., Bródka, P., Bisagno, N., & Conci, N. (2021). PanopTOP: a framework for generating viewpoint-invariant human pose estimation datasets. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 234-242).