Probabilistic 3D Human Shape and Pose Estimation from Multiple Unconstrained Images in the Wild

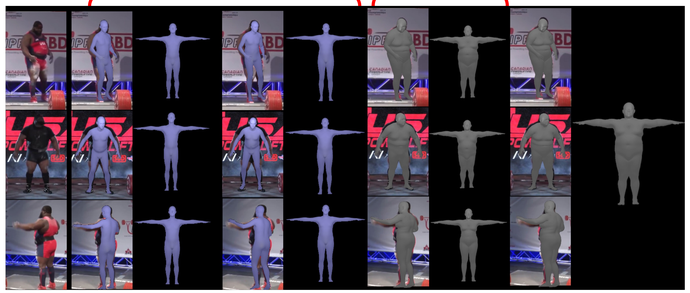
Akash Sengupta Ignas Budvytis Roberto Cipolla

UNIVERSITY OF CAMBRIDGE

CVPR VIRTUAL JUNE 19-25

Introduction

- Aim: Predict **3D body shape and pose** from a ***group* of *unconstrained* images** of a subject.
  - No constraints are imposed on the subject's pose, camera viewpoint or background and lighting conditions between images (unlike video or multi-view methods).
  - We estimate (i) a **single identity-dependent body shape** that is consistent across all images and (ii) a **different body pose** for each image.
- Current approaches provide good pose estimates but **body shapes are inaccurate or inconsistent.**

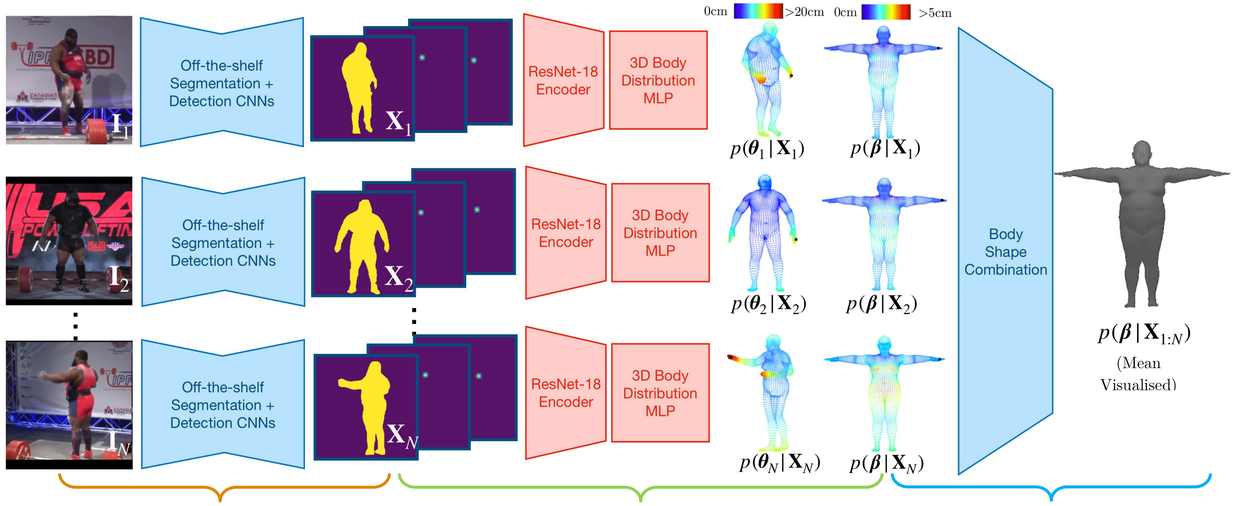


- We predict more **accurate** and **consistent** body shapes by **aggregating the visual shape information present in multiple images** of a subject.

References

1. N. Kolotouros, G. Pavlakos, M. J. Black, and K. Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. ICCV 2019.
2. N. Kolotouros, G. Pavlakos, and K. Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. CVPR 2019.
3. A. Sengupta, I. Budvytis and R. Cipolla. Synthetic Training for Accurate 3D Human Pose and Shape Estimation in the Wild. BMVC 2020.
4. M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: A Skinned Multi-Person Linear model. ACM SIGGRAPH Asia 2015.

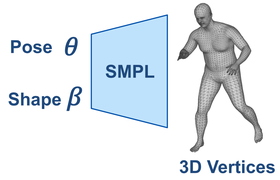
Method



(i) Proxy representation computation

- Group of input images are converted into silhouette + 2D joint heatmap representations are using off-the-shelf CNNs.

Preliminary: SMPL[4]



- Parametric body model mapping mapping **pose (θ)** and **body shape (β)** parameters to 3D vertices.

(ii) Body shape (β) and pose (θ) distribution prediction

- A deep neural network outputs Gaussian distributions over SMPL pose and shape parameters, conditioned on the inputs.

$$p(\boldsymbol{\theta}_n|\mathbf{X}_n) = \mathcal{N}(\boldsymbol{\theta}_n; \boldsymbol{\mu}_\theta(\mathbf{X}_n), \boldsymbol{\Sigma}_\theta(\mathbf{X}_n))$$
$$p(\boldsymbol{\beta}|\mathbf{X}_n) = \mathcal{N}(\boldsymbol{\beta}; \boldsymbol{\mu}_\beta(\mathbf{X}_n), \boldsymbol{\Sigma}_\beta(\mathbf{X}_n))$$

- (Diagonal) covariance matrices represent heteroscedastic aleatoric uncertainty in SMPL parameters, which can arise due to occluded input images.

$$\boldsymbol{\Sigma}_\theta(\mathbf{X}_n) = \mathrm{diag}(\boldsymbol{\sigma}^2_\theta(\mathbf{X}_n))$$
$$\boldsymbol{\Sigma}_\beta(\mathbf{X}_n) = \mathrm{diag}(\boldsymbol{\sigma}^2_\beta(\mathbf{X}_n))$$

Aleatoric Uncertainty

- A negative log-likelihood loss is used for training.

(iii) Probabilistic body shape combination

- Body shape distributions from each input are probabilistically combined into a final distribution.

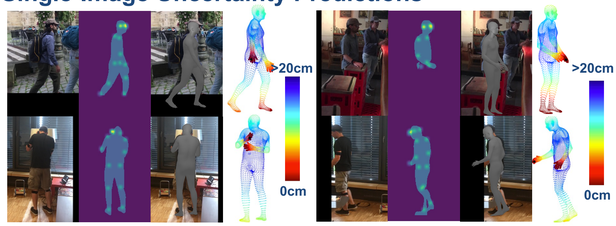
$$p(\boldsymbol{\beta}|\{\mathbf{X}_n\}_{n=1}^N) \propto \prod_{n=1}^{N} p(\boldsymbol{\beta}|\mathbf{X}_n)$$
$$\propto \mathcal{N}(\boldsymbol{\beta}; \mathbf{m}, \mathbf{S})$$

- Combined shape estimate $\mathbf{m}$ is the average of per-input means weighted by the corresponding uncertainties.

$$\mathbf{m} = \mathbf{S}\left(\sum_{n=1}^{N} \boldsymbol{\Sigma}_\beta^{-1}(\mathbf{X}_n)\boldsymbol{\mu}_\beta(\mathbf{X}_n)\right)$$
$$\mathbf{S} = \left(\sum_{n=1}^{N} \boldsymbol{\Sigma}_\beta^{-1}(\mathbf{X}_n)\right)^{-1}$$

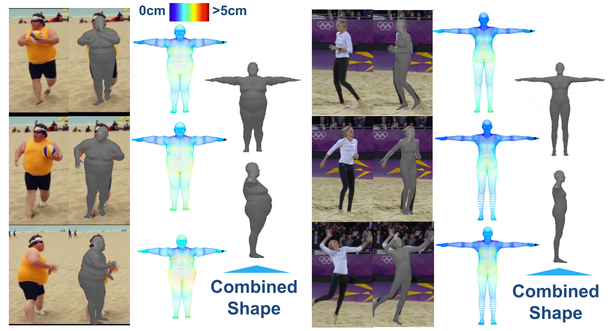
Results

Single-Image Uncertainty Predictions



- Network predicts larger variance (uncertainty) for SMPL parameters corresponding to occluded body parts.

Combined Body Shape Predictions



- Combined shape estimates via uncertainty-weighted averaging are more accurate than individual predictions.

Group size	Method	RMSE (cm) C	S	H	B	F	T
1	SPIN [1]	6.9	8.0	6.6	6.9	2.5	5.3
	STRAPS [3]	6.7	5.3	4.3	3.9	1.8	3.7
	Ours	4.9	4.7	5.5	4.2	1.8	3.9
4	SPIN [1] + Mean	6.5	8.1	6.4	6.7	2.4	5.1
	STRAPS [3] + Mean	6.1	4.2	4.0	**3.2**	1.7	3.3
	Ours + Mean	3.4	3.9	3.8	4.9	**1.6**	3.1
	Ours + PC	**3.1**	**3.8**	**2.7**	5.0	1.7	**2.8**

C = Chest
S = Stomach
H = Hips
B = Biceps
F = Forearms
T = Thighs

- Dataset of tape-measured humans is used to evaluate probabilistic combination (PC) in terms of measurement RMSE.
- PC outperforms current single-image approaches, as well as naïve-averaging ("Mean") of outputs from those approaches.