End-to-end Recovery of Human Shape and Pose

Summary

Angjoo Kanazawa *et al*

CVPR 2018

1 Abstract

- Uses parameterized shape and joint angles in contrast to computing 2D and 3D joint locations
- Minimizes reprojection loss of keypoints
- Infers 3D Pose and shape parameters directly from image pixels
- Uses SMPL which paramaterizes the mesh using joint angles and low dimensional linear space
- Model implicitly learns joint angles from 3D boy model datasets
- Given an image, the conditional generative adversarial network has to infer the 3D mesh parameters and the camera such that the 3D keypoints match the annotated 2D keypoints after projection
- Discriminator network acts as a weak supervision whose task is to determine if the 3D parameters correspond to bodies of real humans

2 Proposed Framework

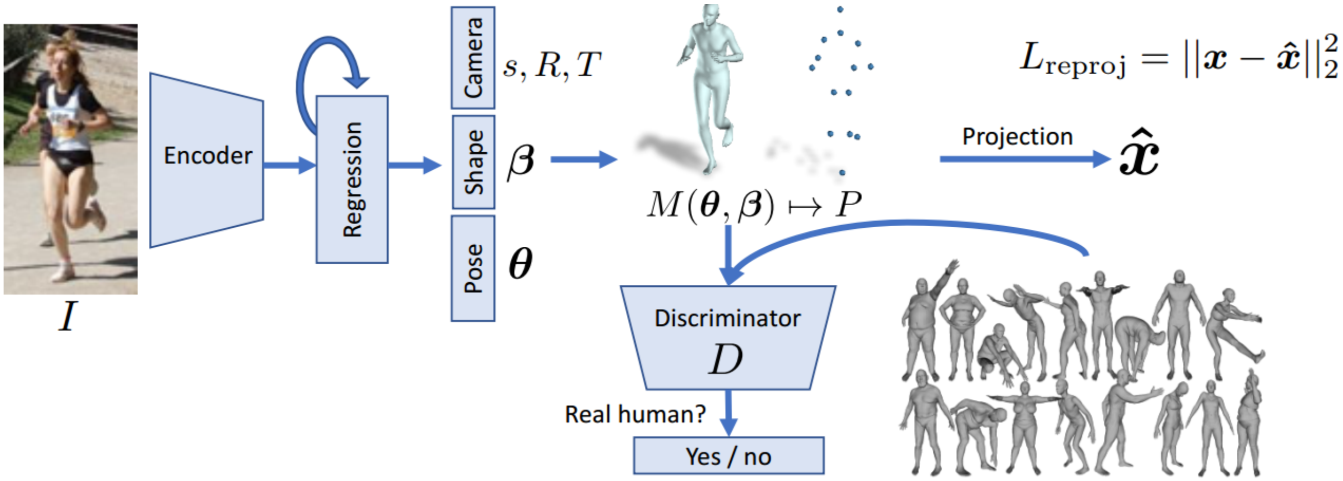

Figure 2: **Overview of the proposed framework.** An image I is passed through a convolutional encoder. This is sent to an iterative 3D regression module that infers the latent 3D representation of the human that minimizes the joint reprojection error. The 3D parameters are also sent to the discriminator D, whose goal is to tell if these parameters come from a real human shape and pose.

3 Method

- Directly infers SMPL parameters from images instead of detected 2D keypoints
- Directly output 3D rotations of joints from images as well as the surface vertices
- Output: Shape as well as the camera scale
- During training all images are annotated with ground truth 2D joints
- Convolutional features of the image are sent to the iterative 3D regression module who infers the 3D human body and the camera such that its 3D joints project onto the annotated 2D joints
- Inferred parameters are also sent to an adversarial discriminator network whose task is to determine if the 3D parameters are real meshes from the unpaired (3D Mesh pool) data
- Overall objective: $\mathrm{L} = \lambda(\mathrm{L}_{\mathrm{reproj}} + 1\mathrm{L}_{\mathrm{3D}}) + \mathrm{L}_{\mathrm{adv}}$
  where λ controls the relative importance of each objective, 1 is an indicator function that is 1 if ground truth 3D is available for an image and 0 otherwise

3.1 Body Representation

- SMPL factors human body into shape and pose.
- Shape $\beta \in \mathrm{R}^{10}$ parameterized by first 10 coefficients of PCA shape space
- Pose $\theta \in \mathrm{R}^{3\mathrm{K}}$ is modeled by relative 3D rotation of K = 23 joints in axis-angle representation

3.2 Iterative 3D Regression with Feedback

- 3D regression module: output Θ given an image encoding ϕ such that the joint reprojection error is minimized: $\mathrm{L}_{\mathrm{reproj}} = \Sigma_i \parallel v_i(x_i - \hat{x}_i) \parallel_1$
- Here $x_i \in R^{\{2\times K\}}$ is the ith ground truth 2D joints and $v_i \in \{0, 1\}^K$ is the visibility (1 if visible, 0 otherwise) for each of the K joints
- 3D Losses:

$$L_{3D} = L_{3Djoints} + L_{smpl}$$
$$L_{joints} = \parallel (X_i - \hat{X}_i) \parallel_2^2$$
$$L_{smpl} = \parallel ([\beta_i, \theta_i] - [\hat{\beta}_i, \hat{\theta}_i]) \parallel_2^2$$

3.3 Factorized Adversarial Prior

- To allow for anthropometrically plausible bodies to be generated, a Discriminator network D is trained to tell whether the SMPL parameters correspond to a plausible real body or not
- Input to each discriminator is very low dimensional (10-D for β, 9-D for each joint and 9K-D for all joints)
- Train K + 2 discriminators. Each discriminator $\mathrm{D_i}$ outputs values between [0, 1], representing the probability that Θ came from the data. Least square formulation used for stability.
- Let$E$ represent the encoder including the image encoder and the 3D module. The adversarial loss function for the encoder:

$$\min L_{adv}(E) = \sum_i \mathbb{E}_{\Theta \sim p_E}[(D_i(E(I)) - 1)^2]$$

- Objective of each discriminator is:

$$\min L(D_i) = \mathbb{E}_{\Theta \sim p_{data}}[(D_i(\Theta) - 1)^2] + \mathbb{E}_{\Theta \sim p_E}[D_i(E(I))^2]$$