End-to-end Recovery of Human Shape and Pose

At this paper the authors present an end-to-end framework for recovering a full 3D mesh of a human body from a single RGB image.

They use the generative human body model, SMPL, which parameterizes the mesh by 3D joint angles and a low dimensional linear shape space. Estimating a 3D mesh opens the door to a wide range of applications such as foreground and part segmentation, which is beyond what is practical with a simple skeleton. The output mesh can be immediately used by animators, modified, measured, manipulated and retargeted. Their output is also holistic – they always infer the full 3D body even in cases of occlusion and truncation.

Overview of the proposed framework- an image 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 re-projection error, the 3D parameters are also sent to the discriminator, whose goal is to tell if these parameters come from a real human shape and pose.

Their approach is similar to 3D interpreter networks in the use of re-projection loss and the more recent adversarial inverse graphics networks for the use of the adversarial prior.

They go beyond the existing techniques in multiple ways:

1. They infer 3D mesh parameters directly from image features, while previous approaches infer them from 2D key points. This avoids the need for two stage training and also avoids throwing away valuable information in the image such as context.

2. Going beyond skeletons, they output meshes, which are more complex and more appropriate for many applications. Again, no additional inference step is needed.

3. Their framework is trained in an end-to-end manner. They out-perform previous approaches that output 3D meshes in terms of 3D joint error and run time.

4. They show results with and without paired 2D-to-3D data. Even without using any paired 2D-to-3D supervision, their approach produces reasonable 3D reconstructions.

3D Body Representation-

They encode the 3D mesh of a human body using the Skinned Multi-Person Linear (SMPL) model. SMPL is a generative model that factors human bodies into shape – how individuals vary in height, weight, body proportions – and pose – how the 3D surface deforms with articulation.

Iterative 3D Regression with Feedback-

The goal of the 3D regression module is to output Θ given an image.

However, directly regressing Θ in one go is a challenging task, particularly because Θ includes rotation parameters. In this work, they take inspiration from previous works and regress Θ in an iterative error feedback (IEF) loop, where progressive changes are made recurrently to the current estimate.

Factorized Adversarial Prior –

The re-projection loss encourages the network to produce a 3D body that explains the 2D joint locations, however anthropometrically implausible 3D bodies or bodies with gross self-intersections may still minimize the re-projection loss. To regularize this, they use a discriminator network that is trained to tell whether SMPL parameters correspond to a real body or not. They refer to this as an adversarial prior as in since the discriminator acts as a data-driven prior that guides the 3D inference.

The results of this paper without using any paired 3D data are promising since they suggest that we can keep on improving our model using more images with 2D labels, which are relatively easy to acquire, instead of ground truth 3D, which is considerably more challenging to acquire in a natural setting.