Master Thesis

# 3D human interaction synthesis for action recognition data augmentation

Anders Bredgaard Thuesen

March 8, 2024

# Abstract

# Introduction

# Data

# Methods

## SMPL & Human Mesh Reconstruction

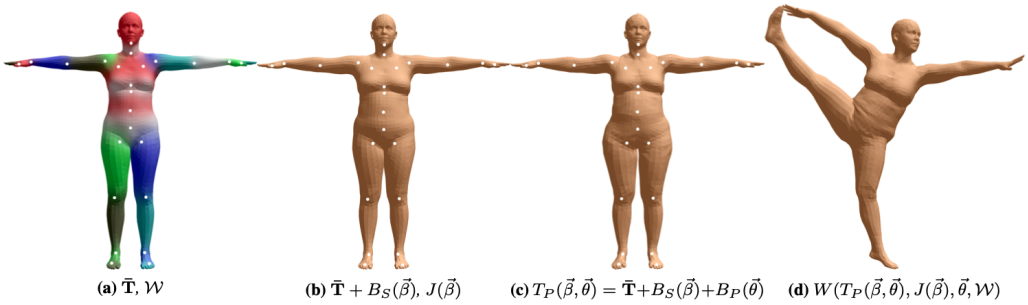


Figure 1: SMPL model

The Skinned Multi-Person Linear (SMPL) model is a parametric body shape model that accurately represents a wide range of human bodies and poses. It is built upon a foundation of linear blend skinning enhanced with corrective blend shapes, which are derived from a large dataset of body scans. The model captures the subtle deformations that occur with different body shapes and poses and can easily be rendered due to its compatibility with existing graphics pipelines. Since its publication, several extensions such as DMPL, incorporating dynamic soft-tissue deformation and SMPL-X, also modelling hands and facial expressions have been introduced. The model is parameterized by $\vec{\beta}$, capturing the variations from a mean body shape and $\vec{\theta}$, specifying the axis-angle rotation of 23 of the template skeleton joints. Mathematically, the model can be expressed as:

$$M(\vec{\beta}, \vec{\theta}) = W(T_P(\vec{\beta}, \vec{\theta}), J(\vec{\beta}), \vec{\theta}, \mathcal{W}) \tag{1}$$

where $T_P(\vec{\beta}, \vec{\theta})$ returns the vertices of the rest pose, incorporating the deformations from the body shape and pose and is given by:

$$T_P(\vec{\beta}, \vec{\theta}) = \bar{\mathbf{T}} + B_S(\vec{\beta}) + B_P(\vec{\theta}) \tag{2}$$

$J(\vec{\beta})$ returns the 3D joint locations from the shaped template vertices using a learned regression matrix $\mathcal{J}$ and is given by:

$$J(\vec{\beta}) = \mathcal{J}(\bar{\mathbf{T}} + B_S(\vec{\beta})) \tag{3}$$

$W$ is the skinning function (e.g. Linear Blend Skinning (LBS) or Dual-Quaternion Blend Skinning (DQBS)) and $\mathcal{W}$ is the blend weights.

## Tracking & Matching

We track the predicted staff and patients across multiple frames, indicated by $t = 1 \dots T$, by iteratively assigning the predictions, $\{P_i^{(t)}\}_{i=1}^{M_t}$, to the latest known tracks, $\{Q_j^{(t)}\}_{j=1}^{N_t}$ by greedily picking the assignment with the minimum cost:

$$\underset{i,j}{\operatorname{argmin}} \, \mathcal{L}_{\text{track}}(P_i^{(t)}, Q_j^{(t-1)}), \tag{4}$$

until either all predictions or latest tracks have been assigned exactly once. In the case of any unassigned predictions, a new track is initialized. After tracks have been assigned, the latest track $Q_j^{(t)}$ is updated with the assigned predictions. We choose the following composite loss function:

$$\mathcal{L}_{\text{track}}(P, Q) = \alpha \left\| P_{\text{3D kpts}} - Q_{\text{3D kpts}} \right\|_2 + \beta \left\| P_{\text{class}} - Q_{\text{class}} \right\|_\infty, \tag{5}$$

incorporating both the Euclidean distance between the 3D joint keypoints as well as the predicted person class (staff/patient), with $\alpha$ and $\beta$ weighting the influence of each term. The tracks, $T_i$, are then greedily matched to the ground truth annotations $G_j$, minimizing the loss:

$$\mathcal{L}_{\text{match}}(T, G) = \sum_{t}^{T} \begin{cases} \left\| a_{\text{track,2D kpts}}^{(t)} - b_{\text{track,2D kpts}}^{(t)} \right\|_2 & \text{if } a_{\text{track}}^{(t)} \in a_{\text{track}} \\ \gamma & \text{otherwise} \end{cases} \tag{6}$$

over the trajectory time horizon $t = 1, 2, \dots, T$ where $\gamma$ is the punishment for not detecting the person.

## Denoising Diffusion Probabilistic Models (DDPM)

1. Forward and backwards process
2. Noise schedule
3. ELBO derivation

## Tranformer

## Rotary Positional Encodings (RoPE)

## Network architecture