Handy: Towards a high fidelity 3D hand shape and appearance model

Rolandos Alexandros Potamias[1] Stylianos Ploumpis[1] Stylianos Moschoglou[1]
Vasileios Triantafyllou[2] Stefanos Zafeiriou[1]
[1]Imperial College London [2]Cosmos Designs Ltd
{r.potamias,s.ploumpis,s.moschoglou,s.zafeiriou}@imperial.ac.uk

A. 3D Hand Reconstruction: Method

The method that we follow to reconstruct 3D hands from single images is composed by three main components. The first module is a ResNet50 network, pretrained on ImageNet, that acts as a feature extractor. Following that, a set of regression branches that predict the latent parameters of the Handy model, i.e. shape, pose and texture are used. Finally, the last module of the proposed method predicts the parameters (scale and translation) of an orthographic camera that is used to render the predicted hand mesh back to the image space. All of the aforementioned branches are composed by an MLP layer and take as input the latent features of ResNet. The full architecture is depicted in Figure 1. We trained the proposed architecture for 250 epochs with Adam optimizer and a learning rate of 5e-5.

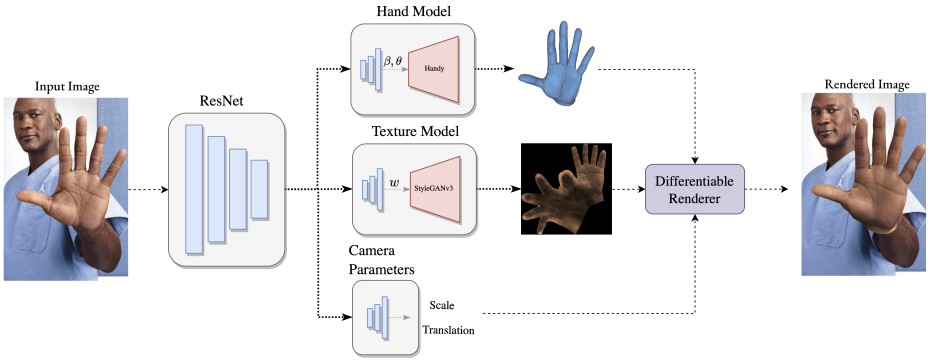

Figure 1. Architecture of the proposed 3d hand reconstruction method.

B. Interpolation on the latent space of the Texture Model

In this section, we showcase the smooth transitions of the latent space of the proposed texture model. In particular, we selected random pairs of UV maps, projected them to the latent space of the texture GAN and then interpolated the values between them. Figure 2 shows that the texture model produces meaningful latent representations between the two UV maps. Additionally, the generated UV maps are feasible and realistic having also very smooth transitions from the source to the target UV maps.

C. Synthetic dataset samples

In this section we visualize several samples of the dataset we used to train the 3D reconstruction module. To generate the sythetic data we used Blender software following [1]. Figure 3 shows examples of the synthetic data used. The composed synthetic dataset has both plain hands (top two rows) and hands interacting with objects (last row).