**Name: Elim Lemango**

**EECS 6322: Project Proposal**

**Paper Name: High-Fidelity Human Avatars from a Single RGB Camera**

**Link:** [**Paper Link**](https://openaccess.thecvf.com/content/CVPR2022/papers/Zhao_High-Fidelity_Human_Avatars_From_a_Single_RGB_Camera_CVPR_2022_paper.pdf)

**Summary**

The study describes a technique for modelling human bodies in 3D depth with just a video from a single RGB camera. The technique uses machine learning techniques to derive the human body's 3D geometry and sharp texture from a video input.

The method first recognises the subject's body parts, such as the arms and legs, and then uses a neural network to forecast the 3D position and orientation of each part. A 3D model of the subject's body is then created using the estimated locations and orientations of the various body parts. In addition, the methodology followed by the researchers, also collects high-resolution texture data from the subject's skin and clothing.

The final 3D model is extremely detailed and precisely reproduces the subject's appearance and body type. The method outperformed existing methods in terms of accuracy and computational efficiency after being tested on two different sets of datasets (public dataset of humans as well as a selfie video dataset collected by the authors). The proposed method has use in industries including virtual reality, gaming, and teleconferencing, according to the paper's conclusion.

**Reproducibility Approach**

The framework that I will be using to reproduce the paper is by using PyTorch. The authors have used two networks in their research which are the dynamic surface network and the reference-based neural rendering network. The losses are given in the supplementary document for the networks are provided in the index section of the paper so I will train the network accordingly. In accordance with the paper, I will train the dynamic surface network first and then train the reference-based neural rendering network.

The main contribution of the paper are as follows, but due to time constraints, I will be working towards reproducing 1 and 3 of the main contributions.

1. *They propose a coarse-to-fine framework which com- bines neural texture with dynamic surface deformation to generate a fully textured avatar from a monocular video captured by the users themselves.*
2. *They propose a dynamic surface network to model the pose-dependent surface deformations of a moving per- son, which deals with the misalignment problem and disentangles the shape and texture of the person.*
3. *They propose a reference-based neural rendering net- work and exploit a bottom-up sharpening-guided fine- tuning strategy, which fuses all the observations into a consistent representation and enables to generate the detailed texture map.*

To this end, I will be evaluating the network, which consists of 24 videos of 11 people, using only the People Snapshot Dataset. Then, I'll continue comparing the results with those of my replicated network, the outcomes of the study, and the two state of the art models that the authors compared their outcomes with (Octopus and Video Avatar).