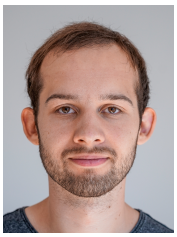
# Learning to Reconstruct 3D Faces by Watching TV

3D Vision Project Proposal
Supervised by: Yao Feng, Weiyang Liu
March 14, 2022


GROUP MEMBERS

Anne Marx Lucas Weitzendorf Deniz Yildiz



## I. DESCRIPTION OF THE PROJECT

The goal of this project is to build a pipeline that improves the generation of person-specific high-quality 3D face models from TV videos. We build face models for each character using a collection of valid facial images from any video.

Reconstructing 3D faces from 2D images is a longstanding yet fundamental problem in computer vision. DECA [1] demonstrates the feasibility of robustly reconstructing a high-fidelity 3D face for a single RGB image without human annotations. However, current pipelines suffer from severe ambiguity and ill-posedness caused by the nature of the problem. To address this, we aim to leverage the abundant information in videos. Particularly, we make use of the fact that main characters frequently show up in TV episodes under a diverse set of environment conditions (lighting, occlusion, etc.), and seek to reconstruct high-fidelity 3D faces from these TV videos. Moreover, this endeavor could be a stepping stone to build a large-scale 3D face dataset that enables better single-view 3D face reconstruction methods.

## II. WORK PACKAGES AND TIMELINE

The first part of the project will focus on the face recognition pipeline. Firstly, we gather training data by downloading episodes of a few shows of our choice. Then, we identify and run the most suitable face detector and recognition algorithms to recover faces from isolated video frames such as SCRFD [2] or YOLO5Face [3]. Next, we will discard results from unsuitable images and cluster the remaining face descriptors to identify relevant recurring characters. During this step, it is important to extract relevant properties for the images needed for the following DECA model. This should be taken into account for the face recognition and clustering models. We aim to finish this within the first five weeks.

The second part of the project will consist of facial reconstruction based on the clustered descriptors. The end goal is to improve on DECA's static image reconstruction by leveraging the implicit temporal data between individual frames. This should be sufficiently challenging to take up the remaining scope of the project.

Given enough time, we also plan on creating a user interface to simplify use of the pipeline. It could also be interesting to extend the pipeline to robustly handle non-professional videos, so it can be used for private face modeling as well.

The language we plan to use is Python. The group members will work together on all of the tasks.

## III. Possible Challenges

One of the challenges will be to incorporate facial images of different qualities, e.g. due to different distances to the camera or fluctuating lighting conditions. It is possible to choose which videos or which specific frames to consider based on their quality, or generalize the method for different resolutions.

A critical part of the detection pipeline will be to identify the similarity threshold for clustering faces. Setting it too high will result characters not being recognized, whereas the opposite will lead to not being able to distinguish between them.

Another challenge could be to determine the best input data for each module in the pipeline. The neural networks have different input and output data dimensions and requirements and might have to be adjusted to fit into the pipeline. For example, DECA models faces from a single image whose properties are dependent on the facial expression. Thus the input image has an impact on the quality of the facial model. This can be tackled by performing various qualitative experiments with each module to retrieve required properties for the module's input to achieve good outputs. This influences the choice for the face recognition model. It might have to be adjusted to output data in regards to these required properties.

## IV. Outcomes and Demonstration

The expected outcome is a working pipeline to generate 3D faces of relevant cast members in any TV show without prior knowledge. We expect to show a pre-recorded demo due to runtime constraints of the detection models.

## References

[1] Feng et al. Learning an animatable detailed 3d face model from in-the-wild images. In *SIGGRAPH*, 2021.

[2] Jia Guo, Jiankang Deng, Alexandros Lattas, and Stefanos Zafeiriou. Sample and computation redistribution for efficient face detection. *arXiv preprint arXiv:2105.04714*, 2021.

[3] Delong Qi, Weijun Tan, Qi Yao, and Jingfeng Liu. Yolo5face: why reinventing a face detector. *arXiv preprint arXiv:2105.12931*, 2021.