Persona: Real-Time Neural 3D Face Reconstruction for Visual Effects on Mobile Devices

Gengdai Liu
BIGO Technology
Singapore
liugengdai@gmail.com

Xiaowei Zhang
BIGO Technology
Singapore
zhangxiaowei1@bigo.sg

Feiqian Zhang
BIGO Technology
Singapore
zhangfeiqian@bigo.sg

Yu Wei
BIGO Technology
Singapore
weiyu@bigo.sg

Figure 1: *Persona* 3D face tracker (left most) and various visual effects built on it.

ABSTRACT

We present *Persona*, a real-time human face-oriented visual effect solution on mobile devices. *Persona* consists of a 3D face tracker with multi-scale reconstruction models for different-level of mobile devices and a visual effect authoring tool. Our face tracker is able to reliably predict a sequence of facial and illumination parameters from a monocular video in real-time. Those parameters can then be used to develop many interesting applications. We demonstrate that our method outperforms existing state-of-the-art work about 3D face reconstruction on mobile devices and showcase results generated by our tool.

CCS CONCEPTS

• **Computing methodologies** → **Computer graphics**; **Mesh geometry models**; *Animation*; **Neural networks**.

KEYWORDS

face reconstruction, expression tracking, neural networks, weakly-supervised learning, mobile device

ACM Reference Format:
Gengdai Liu, Xiaowei Zhang, Feiqian Zhang, and Yu Wei. 2021. Persona: Real-Time Neural 3D Face Reconstruction for Visual Effects on Mobile Devices. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Talks (SIGGRAPH '21 Talks), August 09-13, 2021.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3450623.3464634

SIGGRAPH '21 Talks, August 09-13, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-8373-8/21/08.
https://doi.org/10.1145/3450623.3464634

1 INTRODUCTION

Using a 3D face model as a strong prior to produce visual effects for mobile videos is in great demand. However, accurately predicting 3D face mesh in real-time remains a challenging task. Recently, with the development of Convolutional Neural Networks (CNN), many CNN-based methods directly regress 3D face mesh from videos. Some of these methods train neural networks in a supervised manner using either pre-computed or annotated training data [Grishchenko et al. 2020]. In contrast, [Deng et al. 2019] trains the model in a weakly-supervised fashion with the help of differentiable renderers to avoid the requirement of ground-truth 3D face shape. However, these methods either depend on a 3D face model which fails to predict some common facial expressions or leverage large networks impossible for real-time applications on mobile devices.

To balance accuracy and efficiency, we propose *Persona* which consists of a lightweight face tracker and an authoring tool built on it. In particular, we create our own face model to reliably model various facial expressions, and leverage the idea of knowledge distillation to train the models of different complexities using weakly supervised learning.

2 METHOD

There are two main contributions in *Persona*. Firstly, we create a multi-resolutional 3D Morphable Model (3DMM) which is more diverse than public data sets. Secondly, we train reconstruction models of different complexities in a weakly-supervised way. Moreover, our neural network is able to predict illumination parameters.

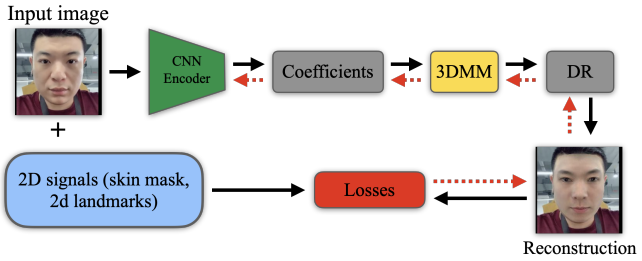


Figure 2: *Persona* model training pipeline. For each face image, we use a CNN encoder to regress 3DMM coefficients, spherical harmonic illumination coefficients, and camera pose. Then, we use these coefficients and our 3D face model to reconstruct 3D face which, together with lighting coefficients and camera pose, is fed to a differentiable renderer (DR) to render face on the input image. The training pipeline only utilizes simple 2D supervision (blue box). Red dashed arrows denote the path of gradient back-propagation from the loss.

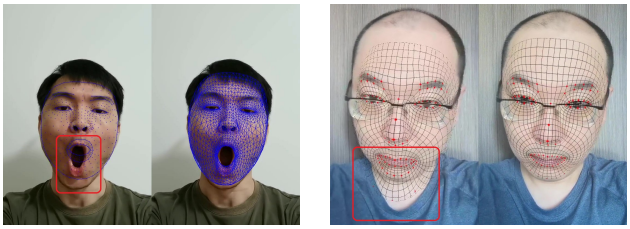
3D face model. We scan faces of 270 identities of various races and ages, each with 20 expressions . We also design a new multi-resolutional face mesh template with up to 20k vertices, and create 46 user-specific blendshapes [Li et al. 2010]. Based on these 3D face meshes, we get a double PCA face model with 80 identity bases, 30 expression bases, and 79 texture bases. To enrich the diversity of our textures, we increase the texture samples by optimizing albedos of 1000 high-quality face images.

Model training: An overview of the model training pipeline is shown in Figure 2. Different from [Deng et al. 2019], we use MobileNetV2, a much smaller network as the encoder, and use stronger losses and more training data to guarantee natural and stable results. For landmark loss, we dynamically select contour 3D landmarks on the mesh, ignore invisible ones due to self-occlusion, and assign adaptive loss weights to different facial regions and head poses. To get accurate mesh for profile, we impose symmetry constraint on mesh for faces with large yaw pose. We also impose box constraints on albedo to avoid over- or under-exposure. Moreover, we elaborately regularize identity and expression coefficients and use more accurate skin masks from dedicated face parsing model. To diversify the training data, we add synthesized images to our training data (see the accompanying video). As a result, we collect more than 350k face images for training. To make our tracker run in real-time on different levels of mobile devices, we adopt the idea of knowledge distillation to train models with different complexities. In particular, we first train a teacher model of high complexity, then train smaller models with less scale and smaller input image size. Eventually, we have three models of high-, mid-, and low-complexity respectively. With our highly optimized inference engine, our models run about 50fps even on low-end devices. See Table 1 for details.

Authoring tool: Persona's authoring tool in our in-house editor consists of four sub-modules: AR makeup, face warping, pin-to-mesh, and lighting estimation. Using the face mesh along with the camera pose predicted by our tracker to create AR makeup is straightforward. Designers can use any textures, lighting and shaders on the mesh to produce makeup effects they expect. The face warping module can warp a user's face by deforming the user's face mesh and then rendering it over the original face in the video frame. The deformation is transferred from a rigged base mesh by simply adding the vertex displacements of the base mesh to the user's face mesh. The base mesh can be rigged using any method, e.g. skeleton, blendshape and etc. In the case of extreme deformation, the background warping is performed to eliminate the gap on the boundary between the rendered user face and the video frame. The pin-to-mesh module allows designers to attach an object to a point on the face mesh in the UV space. With the UV coordinates, the barycentric coordinates and normal vector of the point are calculated, and then the attached object follows the mesh during deformations. In the lighting estimation module, the illumination coefficients output from the tracker are converted into the lighting intensity and direction, so that the face mesh is used as a probe to detect the lighting of the environment. Hence, the AR objects can be seamlessly merged into the video with the shading using the estimated lighting.

Table 1: Running time of our face tracker on different chips.

	Snapdragon 865	Snapdragon 665	Snapdragon 425
High	**5.4 ms**	19.5ms	55.3ms
Mid	2.8 ms	**11.5ms**	30.7ms
Low	2.2 ms	8.8ms	**22.6ms**



(a) Persona vs MediaPipe **(b) Persona vs SparkAR**

Figure 3: Comparing Persona face tracker (on the right of each figure) with MediaPipe (a) and SparkAR (b).

3 RESULTS

We have compared our tracker with some state-of-the-art work including Google's MediaPipe [Grishchenko et al. 2020] and Facebook's Spark AR [1]. Figure 3 shows that our method outperforms these existing work. With our stable and accurate tracker and the authoring tool, Persona has been used in the production of many interesting visual effects, as shown in Figure 1 and the accompanying video. We are still seeking to extend it for more creative effects.

REFERENCES

Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. 2019. Accurate 3D Face Reconstruction With Weakly-Supervised Learning: From Single Image to Image Set. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops.*

Ivan Grishchenko, Artsiom Ablavatski, Yury Kartynnik, Karthik Raveendran, and Matthias Grundmann. 2020. Attention Mesh: High-fidelity Face Mesh Prediction in Real-time. arXiv:2006.10962 [cs.CV]

Hao Li, Thibaut Weise, and Mark Pauly. 2010. Example-Based Facial Rigging. *ACM Transactions on Graphics (Proceedings SIGGRAPH 2010)* 29, 3 (July 2010).

[1] https://sparkar.facebook.com/ar-studio/