# FaceMagic: Real-time Facial Detail Effects on Mobile

Llogari Casas
Disney Research Los Angeles,
Edinburgh Napier University &
3FINERY LTD

Yue Li
Disney Research Los Angeles &
University of Pennsylvania

Kenny Mitchell
Disney Research Los Angeles,
Edinburgh Napier University &
3FINERY LTD


**Figure 1: Guests using our *Face Magic* framework with themed detailed steel stylized face filters. Our system performs camera ingest to coarse mesh depth generation in .43ms, face depth smoothing in .52ms, patch based psuedo-inverse local reconstruction in 12.22ms, and our detail filter effects shaders in 1ms, leading to a total real-time processing in 15ms per frame.**

## ABSTRACT

We present a novel real-time face detail reconstruction method capable of recovering high quality geometry on consumer mobile devices. Our system firstly uses a morphable model and semantic segmentation of facial parts to achieve robust self-calibration. We then capture fine-scale surface details using a patch-based *Shape from Shading (SfS)* approach. We pre-compute the patch-wise constant Moore–Penrose inverse matrix of the resulting linear system to achieve real-time performance. Our method achieves high interactive frame-rates and experiments show that our new approach is capable of reconstructing high-fidelity geometry with corresponding results to off-line techniques. We illustrate this through a variety of comparisons with off-line and on-line related works, and include demonstrations of novel face detail shader effects processing.

## CCS CONCEPTS

• **Computing methodologies** → **Mixed / Augmented Reality**.

**ACM Reference Format:**
Llogari Casas, Yue Li, and Kenny Mitchell. 2020. FaceMagic: Real-time Facial Detail Effects on Mobile. In *SIGGRAPH Asia 2020 Technical Communications (SA '20 Technical Communications), December 4–13, 2020, Virtual Event, Republic of Korea.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3410700.3425429

https://doi.org/10.1145/3410700.3425429

## 1 INTRODUCTION

Recreating plausible and realistic human faces in any *Mixed Reality (MR)* environment plays a key role in conveying identity, message, emotion and intent. For this reason, the computer graphics and vision communities started very early in building computerized tools for analysing real-world faces with the goal of generating digital face images. To reconstruct 3D face models from 2D images, *Shape from Shading (SfS)* is commonly used. This technique estimates the surface normals of objects by observing those objects under different lighting conditions. This approach is grounded on the basis that the amount of light reflected by a surface is dependent on the orientation of the surface in relation to the light source and the observer.

State-of-the-art solutions are capable of producing movie-quality photo-realistic results. However, these systems tend to be bulky, expensive and generally require taking multiple shots. In this research, we introduce an algorithm capable of preserving detailed 3D face reconstructions in real-time using a conventional mobile device with an alike fidelity. The following contributions are introduced:

- we present a novel real-time face detail reconstruction method capable of recovering pixel accurate face geometry on consumer mobile devices.
- we extend from Li et al. [2018], an off-line feature-preserving detailed 3D face reconstruction method, to capture fine-scale surface details using a SfS approach in real-time. We achieved this by dividing the face region into patches for parallelization and pre-computing the patch-wise constant data.
- we demonstrate novel facial effects processing applications derived from high quality face geometry reconstruction with pixel accuracy and show that our results are comparable to off-line techniques.

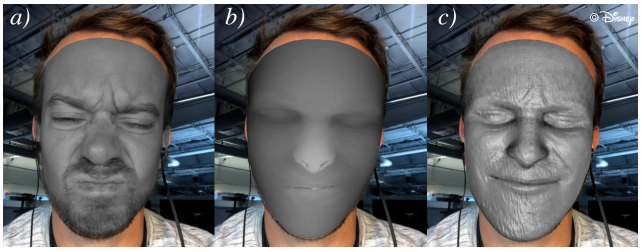


**Figure 2: *a)* is the image luminance map (Y). *b)* is the global facial depth of the coarse mesh without detail. *c)* is the detailed depth map with fine pixel scale geometry, such as wrinkles, that our algorithm solves for (presented with a virtual light source diffuse shader to highlight details).**

## 2 RELATED WORK

Section 2.1 describes related work on prior-aided face reconstruction approaches. Section 2.2 draws from previous on-line approaches to capture fine-scale details using a SfS approach in real-time.

### 2.1 Prior Guided Face Reconstruction

Early research in this topic approached face reconstruction with prior guidance of face shape, expressions and albedo. Blanz and Vetter [2003] presents, *3D Morphable Model (3DMM)*, a parameteric face geometry and albedo model obtained from performing *Principle Component Analysis (PCA)* on a set of registered 3D faces from scanned data.

State-of-the-art techniques mainly focus on *Machine Learning (ML)* approaches. Cao et al. [2014] built a large dataset of identities and expressions by learning the high-order relationship between them. Booth et al. [2016] created a large-scale 3DMM dataset from 10,000 scanned faces from a large variety of the human population. Li et al. [2017] used the skinned model from Loper et al.[2015] and learned the reconstructed model from both static and dynamic scans. Ranjan et al. [2018] used a *Graph Convolution Network (GCN)* to learn the parametric model from the dataset of Li et al. [2017]. Prior-aided approaches are generally helpful to obtain a rough estimation of the face geometry and albedo, but are not accurate enough to model the facial features that define an identity of a person. Our approach uses a pre-defined low-geometry mesh object that represents high level topology for the face, which conforms a generic face model to match the dimensions, shape, and current expression of the detected face. We use this base model as the foundation for overlaying content that follows the user's face shape.

Fitting 3DMM coefficients and adding a displacement afterwards can produce large silhouettes and component matching errors due to the low-dimensitionality of the parametric model. Luo et al. [2017] embedded a medium layer for shape correction, but this approach was still prone to imperfections and carried artifacts. Li et al.[2018] proposed to recover high frequency geometry details by utilizing an albedo prior masks during SFS process for better removal of highlights and self-shadows. They also include a fast proportionality constraint between depth and image gradients to approximate local self-occlusion behavior. In this work, we extend this method to achieve high-fidelity geometry with a patch-based method that includes details such as wrinkles in real-time on mobile GPUs.

### 2.2 Real-time Face Reconstruction

On-line face reconstruction captures the dynamic 3D model of a human face in real-time. Weise et al. [2011] developed a system for on-line 3D face reconstruction by building a dataset from scanned models and then finding the optimal coefficient in real-time. Bouaziz et al. [2013] proposed a solution to avoid the tedious scanning process and dynamically builds the user's blend-shape on-line. Previous work [Cao et al. 2018, 2014] presented solutions to real-time 3D face reconstruction based on a single RGB image and further solves the rigid stabilization problem. Shi et al. [2014] and Garrido et al. [2016] used SfS to reconstruct facial details from video sequences. Cao et al.[2015] learned mapping from texture maps to mesh objects from high-quality scans. Guo et al. [2019] performed reconstruction in real-time but with less accurate silhouette matching. While deep learning approaches require careful training design and appropriate datasets, analytical methods can generally be applied without dependence on training data and directly realize the observed principles, and therefore fall within the focus of this work.

## 3 METHOD

Our framework capable of recovering high quality pixel accurate face geometry in consumer mobile devices. Section 3.1 introduces the approach that our framework uses to first segment the area of the image in which the reconstructed face should be rendered and further how we conform to each individual's facial expressions. Section 3.2 describes how we retrieve the luminance map of the user's face. Section 3.3 details how the input image is used to create a non-detail depth estimation of the segmented area. Section 3.4 presents our feature preserving 3D face reconstruction solver.

### 3.1 Face Segmentation & Expression Conformance

In order to segment the guest's face in the image, we make use of ARKit [2019]. Concretely, we perform an AR face tracking session to obtain a virtual anchor of the user's face location. Doing so, we can automatically retrieve the position and rotation of the face anchor when it gets detected in the camera video feed (see figure 2a). The inherited transform property describes the face's current position and orientation in world coordinates. This transform matrix creates a face coordinates system for positioning additional overlaid elements relative to the face.

Once the virtual anchor is detected, our approach uses a predefined low-geometry mesh object that represents high level topology for the face, which conforms a generic face model to match the dimensions, shape, and current expression of the detected face. We use this base model as the basis for overlaying content that follows the user's face shape. Further, we apply blend-shape coefficient weight activations to retrieve a high-level model of the current facial expression, which is described via a series of many named poses that represent the movement of specific facial features relative to their neutral configurations.

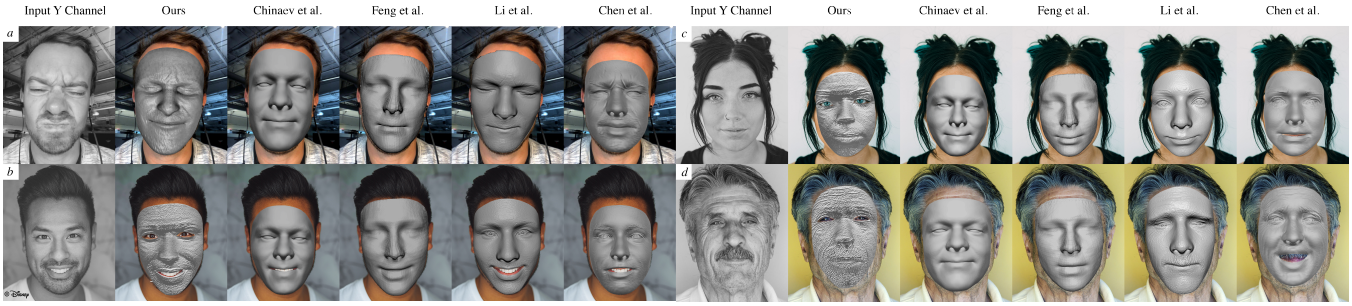


**Figure 3: Results using our real-time face detail reconstruction solver for different subjects. Our algorithm bases its displacement on the input Y luminance map. We compare to Chinaev et al. [2018] for on-line reconstruction in mobile phone and to Li et al. [2018], Feng et al.[2018] and Chen et al. [2019] for off-line high quality techniques.**

## 3.2 Facial Luminance Map

ARKit [2019] persistently captures video frames from the device camera at the application frame rate. It captures pixel buffers in a full-range planar YCbCr format according to the $ITUR.601 - 4$ standard. For greatest performance, we incur the bandwidth cost of the luminance map Y channel only, to preserve detailed features when performing 3D face reconstruction in real-time (see figure 2a). Similar to Beeler et al. [2010] we use the gradients on the luminance map to enhance geometry details as described in section 3.4 to accurately reconstruct fine geometry details, such as wrinkles, on a texel level.

## 3.3 Global Facial Depth Map

Using the pre-defined low geometry density mesh object that represents high level topology for the face, we can then obtain a global depth map of the user's face (see figure 2b). This allows us to create an initial depth map of the mesh without wrinkle-level detail on it. As this mesh is sparse with a low vertex count of 1220 vertices, a direct use of the depth map would exhibit faceting. Therefore, we apply a gradient aware depth map smoothing pass in less than half a millisecond. This depth map is used on the algorithm described in section 3.4 to accurately determine the global position of each texel on the user's face. This approach avoids false positive detail displacements over the face that would result in undesired artifacts. Equation 1 describes the use of the global facial depth map in our formulation.

## 3.4 Patch-based Feature Preserving Solver

We minimize the following energy for the desired depth map,

$$\underset{D^*}{\operatorname{argmin}}\{||D^* - D||^2 + ||\nabla D^* - \nabla L||^2 + ||\Delta D^*||^2\} \tag{1}$$

where $L$ is the image luminance map (see figure 2a), $D$ is the initial depth map of the mesh without details (see figure 2b) and $D^*$ is the depth map with fine geometry details that our algorithm solves for (see figure 2c).

We optimize this problem on a mobile GPU by dividing the input image ($\mathcal{I}$) into a set of squared patches ($\mathcal{P}$) and solve the displacements locally within these patches. This results in solving linear systems of equations taking the form of equation 2 in every patch over the face region per frame. The first term regularizes the solution to not deviate too much from its initial position and the Laplacian terms ensures spacial smoothness. The second term requires the gradiant of the target depth map to be similar to the gradient from the luminance map.

$$Ax = b \tag{2}$$

In order to achieve real-time performance, we pick the size ($p$) of the patches to be 8 in our implemenation. Note that the coefficient matrix ($A$) is constant for each patch, thus the pseudo-inverse could be pre-computed and remains constant. We pre-compute the Moore–Penrose [1920] pseudo-inverse ($A^*$) of ($A$) following equation 3. and source it once only to the fragment shader.

$$A^* = (A^T A)^{-1} A^T \tag{3}$$

In the fragment shader, for each pixel ($p_i$) that the GPU rasterizes, we construct the linear systems from all pixels within the patch it lies in.

We compute equation 4 in the fragment shader. ($x$) is a vector contains the solved depth value for all neighboring pixels inside a patch ($\mathcal{P}_i$). We only fetch the value of ($x_i$) for the pixel ($p_i$) that the GPU is currently rasterizing.

$$x = A^* * b \tag{4}$$

# 4 PERFORMANCE

Our method achieves real-time performance in low powered mobile devices with a queued series of processing stages on the mobile GPU, releasing the CPU for other application processing. In our implementation, we have profiled the performance of our method in an Apple iPhone X. In this device, we achieve a steady rendering rate of 60 frames per second, where an average frame renderers in under 15ms. In detail, our method takes 0.43ms for mesh depth generation, 0.52ms for smoothing, 12.22ms for our patch based reconstruction and 1ms for the final effects composition pass (figure 4). Mobile shaders run directly using Metal Shading Language, which leverages the full computing power of an Apple A11, 3 core, Graphics Processing Unit (GPU) and achieves a performance gain of 24% compared to OpenGL ES 2.

While in our implementation the patch size is adjusted to obtain the maximum detail while preserving 60 frames per second, this could be adjusted according to the device's capability. Therefore, given that mobile computing power is constantly increasing, we understand that our method will be capable of achieving increased reconstructed details with higher resolution multi-camera devices.

## 5 CONCLUSIONS & LIMITATIONS

In this paper, we have introduced a novel real-time facial detail reconstruction method capable of retrieving high-quality facial geometry with pixel precision on consumer mobile devices. While granularity appears in our results due to reconstruction noise inherited from patch segmentation, we still achieve high fidelity wrinkle level details in real time on mobile devices. Traditionally, SfS has been implemented on high-end computing systems, and this work has for the first time introduced a real-time analytical approach that is capable of using SfS in low-powered devices. With advances in mobile computing power, we foresee the possibility to reduce granularity by extending the patch size and adding further post-processing steps to refine our reconstructions.

While our method performs well on a variety of subjects with nearly instantaneous results, for larger scale wrinkles such as can be seen in figure 3.d) both the coarse mesh reconstruction and fine patch can miss these mid-level details. For figure 3.a), a comparison between the reconstructed mesh of Chen et al. [2019] and ours achieves a *Peak Signal-to-Noise Ratio (PSNR)* value of 17.64dB and a *Structural Similarity Index Measure (SSIM)* of 0.81. Therefore, the reconstructed details are equivalent to the off-line results, with local aliasing being the main difference given the lower resolution of the camera sensor when executed in real-time introduces these artefacts. While the reconstructed details achieve temporal consistency for facial wrinkles and features, further temporal methods could address noise and aliasing related artefacts. Although this is an undesired artefact, it is directly related to the computational power of the device and can be reduced with advancements on mobile computing power for higher image resolution processing in real-time hand-held AR.

We compare to Li et al.[2018] and Deng et al. [2019] with the level of detail that our results obtain, while still preserving a low-latency high frame rate on a mobile device (figure 3). While Chinaev et al. [2018] has real-time capabilities, the computed results lack of expression detail resulting in a flat face geometry appearance. We foresee, further sub-pixel detail reconstruction arising from learned detail inference and super resolution methods alongside improvements to camera sensor quality and resolution.

**Figure 4: A tarnished glass facial detail effect in a room with yellow lanterns. Our reconstructed detail normal map shown on the left with an RGB color mapping provides the local surface normal to refract the per pixel view vector into a captured environment cube map, as shown in the middle. A further glass shader variation is shown on the right.**

## REFERENCES


ARKit. 2019. Apple ARKit. https://developer.apple.com/arkit/

Thabo Beeler, Bernd Bickel, Paul Beardsley, Bob Sumner, and Markus Gross. 2010. High-quality Single-shot Capture of Facial Geometry. *ACM Trans. Graph.* 29, 4, Article 40 (July 2010), 9 pages.

Volker Blanz and Thomas Vetter. 2003. Face recognition based on fitting a 3D morphable model. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 25, 9 (2003).

James Booth, Anastasios Roussos, Stefanos Zafeiriou, Allan Ponniah, and David Dunaway. 2016. A 3D Morphable Model Learnt from 10,000 Faces. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5543–5552.

Sofien Bouaziz, Yangang Wang, and Mark Pauly. 2013. Online Modeling for Realtime Facial Animation. *ACM Trans. Graph.* 32, 4, Article 40 (July 2013), 10 pages.

Chen Cao, Derek Bradley, Kun Zhou, and Thabo Beeler. 2015. Real-time High-fidelity Facial Performance Capture. *ACM Trans. Graph.* 34, 4, Article 46 (July 2015), 9 pages.

Chen Cao, Menglei Chai, Oliver Woodford, and Linjie Luo. 2018. Stabilized Real-time Face Tracking via a Learned Dynamic Rigidity Prior. *ACM Trans. Graph.* 37, 6, Article 233 (Dec. 2018), 11 pages.

Chen Cao, Qiming Hou, and Kun Zhou. 2014. Displaced Dynamic Expression Regression for Real-time Facial Tracking and Animation. *ACM Trans. Graph.* 33, 4, Article 43 (July 2014), 10 pages.

Anpei Chen, Zhang Chen, Guli Zhang, Kenny Mitchell, and Jingyi Yu. 2019. Photo-Realistic Facial Details Synthesis from Single Image. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 9429–9439.

Nikolai Chinaev, Alexander Chigorin, and Ivan Laptev. 2018. MobileFace: 3D Face Reconstruction with Efficient CNN Regression. *European Conference on Computer Vision (ECCV) Workshops*.

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 *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 285–295.

Pablo Garrido, Michael Zollhöfer, Dan Casas, Levi Valgaerts, Kiran Varanasi, Patrick Pérez, and Christian Theobalt. 2016. Reconstruction of Personalized 3D Face Rigs from Monocular Video. *ACM Trans. Graph.* 35, 3, Article 28 (May 2016), 15 pages.

Yudong Guo, Juyong Zhang, Jianfei Cai, Boyi Jiang, and Jianmin Zheng. 2019. CNN-Based Real-Time Dense Face Reconstruction with Inverse-Rendered Photo-Realistic Face Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 6 (2019), 1294–1307.

Luo Jiang, Juyong Zhang, Bailin Deng, Hao Li, and Ligang Liu. 2017. 3D Face Reconstruction with Geometry Details from a Single Image. *IEEE Transactions on Image Processing* PP (02 2017).

Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. 2017. Learning a Model of Facial Shape and Expression from 4D Scans. *ACM Trans. Graph.* 36, 6, Article 194 (Nov. 2017), 17 pages.

Yue Li, Liqian Ma, Haoqiang Fan, and Kenny Mitchell. 2018. Feature-preserving Detailed 3D Face Reconstruction from a Single Image. In *Proceedings of the 15th ACM SIGGRAPH European Conference on Visual Media Production (CVMP '18)*.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: A Skinned Multi-person Linear Model. *ACM Trans. Graph.* 34, 6, Article 248 (Oct. 2015), 16 pages.

Eliakim Hastings Moore. 1920. On the reciprocal of the general algebraic matrix. *Bull. Amer. Math. Soc.* 26, 9 (06 1920), 394–395.

Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J. Black. 2018. Generating 3D Faces using Convolutional Mesh Autoencoders. In *European Conference on Computer Vision (ECCV)*, Vol. Lecture Notes in Computer Science, vol 11207. Springer, Cham, 725–741.

Fuhao Shi, Hsiang-Tao Wu, Xin Tong, and Jinxiang Chai. 2014. Automatic Acquisition of High-fidelity Facial Performances Using Monocular Videos. *ACM Trans. Graph.* 33, 6, Article 222 (Nov. 2014), 13 pages.

Thibaut Weise, Sofien Bouaziz, Hao Li, and Mark Pauly. 2011. Realtime Performance-based Facial Animation. In *ACM SIGGRAPH 2011 Papers (SIGGRAPH '11)*. ACM, New York, NY, USA, Article 77, 10 pages.