Real-time Face Video Swapping From A Single Portrait

Luming Ma
University of Houston
lma15@uh.edu

Zhigang Deng
University of Houston
zdeng4@uh.edu

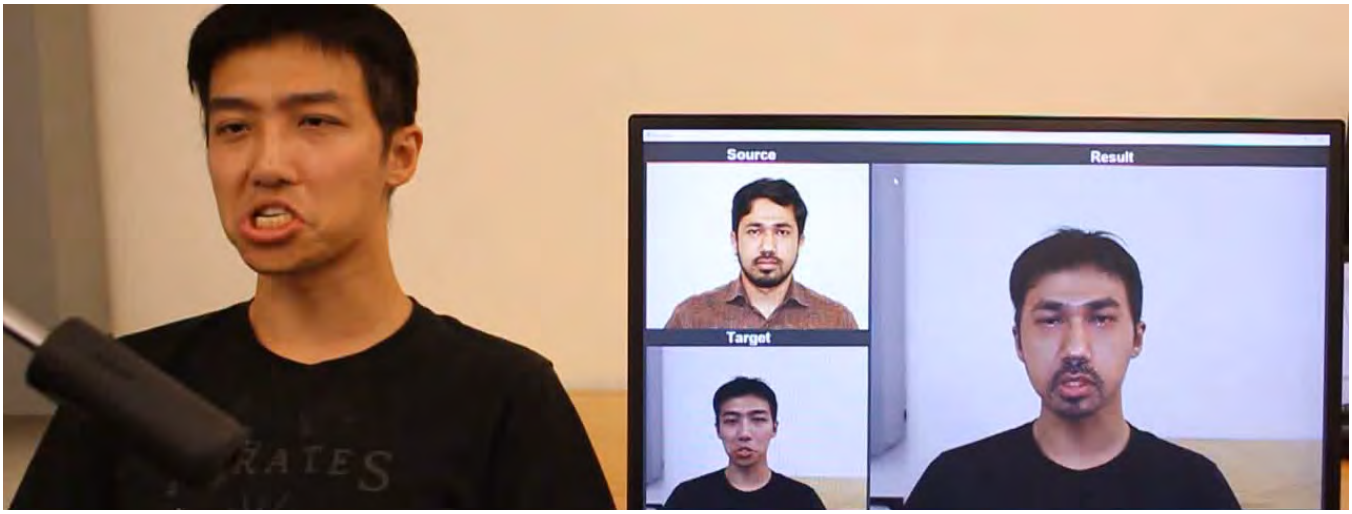


Figure 1: Our system swaps the face from a single source portrait image into an RGB live video stream. The result video retains the facial performance of the target actor while with the identity of the source.

ABSTRACT

We present a novel high-fidelity real-time method to replace the face in a target video clip by the face from a *single* source portrait image. Specifically, we first reconstruct the illumination, albedo, camera parameters, and wrinkle-level geometric details from both the source image and the target video. Then, the albedo of the source face is modified by a novel harmonization method to match the target face. Finally, the source face is re-rendered and blended into the target video using the lighting and camera parameters from the target video. Our method runs fully automatically and at real-time rate on any target face captured by cameras or from legacy video. More importantly, unlike existing deep learning based methods, our method does not need to pre-train *any* models, i.e., pre-collecting a large image/video dataset of the source or target face for model training is not needed. We demonstrate that a high level of video-realism can be achieved by our method on a variety of human faces with different identities, ethnicities, skin colors, and expressions.

I3D '20, May 5–7, 2020, San Francisco, CA, USA

ACM ISBN 978-1-4503-7589-4/20/05…$15.00
https://doi.org/10.1145/3384382.3384519

CCS CONCEPTS

• **Computing methodologies** → **Shape modeling**; *Image-based rendering*;

KEYWORDS

Real-time 3D face reconstruction, face swapping, photorealistic rendering

ACM Reference Format:
Luming Ma and Zhigang Deng. 2020. Real-time Face Video Swapping From A Single Portrait. In *Symposium on Interactive 3D Graphics and Games (I3D '20), May 5–7, 2020, San Francisco, CA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3384382.3384519

1 INTRODUCTION

Face swapping or replacement has been a very active research field in recent years. One of typical face swapping scenarios can be described as follows: given a target video/image, the appearance of the inner face is swapped by the face from a source video/image, while the facial expression, skin color, hair, illumination, and background of the target video/image are preserved [Dale et al. 2011; Garrido et al. 2014]. To date, a number of off-the-shelf applications have been designed to achieve this goal, including Deepfakes [Deepfakes 2019] and Face Swap[1].

Despite potential legal and ethical concerns have emerged in the society in recent years, the face swapping technique itself has rich research values and numerous useful application scenarios in film taking, video editing, and identity protection. For instance, the face

[1]https://faceswap.ms/

of a stunt actor who performs in a dangerous environment can be replaced by a star actor's face captured in a safe studio. It is also applicable to revive the dead actors in legacy films by replacing with the face of the substitute. For video amateurs, an automatic tool that can put the faces of themselves or friends into movie or video clips to create fun content with minimal manual involvement is in great demanding. Furthermore, replacing the face with another identity or virtual avatar in real-time video streaming or conference could be practically needed to protect identity privacy.

Even though noticeable progresses have been made on face swapping over the past several years, *video-realistic* face swapping is still challenging. The differences of face shapes, expressions, head poses, and illuminations between the source and the target faces have posed significant difficulties on the problem. In addition, the human eyes are particularly sensitive to the imperfection in synthesized facial performance and appearance. Previously, researchers sought to tackle the problem by searching for the most similar images/frames from an image database [Bitouk et al. 2008] or video frames [Garrido et al. 2014] and replacing faces through image warping. This line of methods highly relies on the similarity of head poses, expressions, and illuminations between the source and the target images. Another line of approaches resorted to reconstruct 3D face models from both the source and the target images and then re-render the source face into the target background photo-realistically. Although promising results have been presented [Blanz et al. 2004; Dale et al. 2011], these methods typically involve manual interventions (e.g., face alignment) from users. More recently, deep learning approaches [Deepfakes 2019] have been proposed to automatically swap the faces of two identities. However, they require a large image dataset of the face identities and expensive training of the model before running, which undermines the wide applicability, accessibility, and generality of these methods.

In this paper, we propose a new, automatic, real-time method to swap the face in the target video by the face from a *single* source portrait image. Just imagine a selfie image of yourself and an actor interview video clip are given, our method can create a new video clip in which you were taking the interview. In our method, 3D face models with wrinkle level details, appearances, head poses, and illuminations are first reconstructed from the source image and the target video, respectively. Then, a novel face image is rendered using the identity, predicted wrinkles and adapted albedo of the source face and the head pose, expression and illumination of the target face.

Compared to state of the art methods, the main advantages of our method include: (i) little dependency of the source face data (i.e., only need a single still portrait image), (ii) fully automatic and real-time processing, and (iii) swapping both face shape and appearance. More importantly, unlike existing deep learning based methods, our method does not require any assumption of the input face nor require any training data; therefore, our method does not need to collect a large amount of face images for expensive and time-consuming model training, which can bring significant convenience and efficiency to users. As a result, our method can also generalize well to unseen faces.

In sum, the contributions of this work include:

- an automatic real-time system to swap the face in a monocular RGB video by the face from a single portrait image;
- a method to predict wrinkle dynamics of the source face in target expressions; and
- an appearance harmonization method to video-realistically blend the synthesized face into the target video.

2 RELATED WORK

Our method consists of 3D face tracking and swapping from video. We briefly describe recent related works in the two directions below.

3D Face Tracking. Reconstruction of 3D face from video is crucial in graphics and enables many applications in games, films and VR/AR [Zollhöfer et al. 2018]. A significant body of works stem from the seminal morphable face model [Blanz and Vetter 1999], where a statistical model is learned from face scans and later employed to reconstruct facial identity and expression from videos or images. Similarly, [Vlasic et al. 2005] and [Cao et al. 2014b] use multi-linear face models to capture large scale facial expressions. Due to the strong data prior constraint, they cannot capture high frequency details such as wrinkles, which requires further refinements [Bermano et al. 2014]. To reconstruct high resolution face models, structured light and photometric stereo methods [Ma et al. 2008; Zhang et al. 2004] were proposed for face scanning. Passive solutions using multi-view images [Beeler et al. 2010, 2012, 2011; Gotardo et al. 2018] and binocular cameras [Valgaerts et al. 2012] are capable of capturing pore-level geometric details. However, the aforementioned methods usually require delicate camera and lighting setup in controlled environment, which is unfriendly to amateur users and also lacks the ability to process online video clips. In recent years, monocular methods [Fyffe et al. 2014; Garrido et al. 2013; Shi et al. 2014; Suwajanakorn et al. 2014] have shown that shape-from-shading techniques [Horn 1975] can obtain fine-scale details from single RGB video, which opens up the door to build 3D faces from legacy video. The works of [Garrido et al. 2016; Ichim et al. 2015; Suwajanakorn et al. 2015] can build fully rigged face models and appearance from monocular video. All the above methods, however, require intensive off-line processing and are not applicable to real-time applications.

Real-time face capture methods were first developed using RGB-D cameras [Chen et al. 2013; Hsieh et al. 2015; Li et al. 2013; Weise et al. 2011; Zollhöfer et al. 2014]. Later, Cao et al. [2014a; 2013] proposed to capture coarse geometry using regression based face tracking from a RGB camera. Their follow-up work [Cao et al. 2015] learns displacement patches from 2D images to predict medium-scale details. Recently Ma and Deng [2019b] proposed a hierarchical method to capture wrinkle-level face model via vertex displacement optimization on GPU in real-time. Another category of methods solves the 3D face reconstruction problem using deep learning techniques, including CNN [Guo et al. 2018; Sela et al. 2017; Tewari et al. 2018] and autoencoder [Bagautdinov et al. 2018; Lombardi et al. 2018; Tewari et al. 2017; Wu et al. 2018].

Face Reenactment. Face Reenactment transfers the expression of a source actor to a target video. Researchers proposed to use a RGB-D camera to transfer facial expressions in real-time [Thies et al. 2015, 2016; Xu et al. 2014]. Useful scenarios of this technique include Vdub [Garrido et al. 2015], which transfers a dubber's mouth

motion to the actor in the target video; FaceVR [Thies et al. 2018a] which transfers the facial expression of a source actor who is wearing a head-mounted display (HMD) to the target video; and portrait animation which transfers the source expression to a portrait image [Averbuch-Elor et al. 2017] or video [Thies et al. 2018b]. [Ma and Deng 2019a] directly reenacts the facial expression from video in real-time without the driving actor by learning expression correlations using a deep learning approach.

Face Swapping. Most face swapping methods can be categorized into image-based, model-based, and learning based. *2D image-based* methods [Garrido et al. 2014] select the most similar frame from the source video and warp it to the target face. Image-to-image methods [Bitouk et al. 2008; Kemelmacher-Shlizerman 2016] swap the face by automatically selecting the closest face from a large face database. Even though compelling results are produced, they cannot be applied to video since the temporal consistency is not considered. *3D model-based* methods [Blanz et al. 2004; Dale et al. 2011] track the facial performance for both the source and the target faces and re-render the source face under target conditions. Our method is also model-based, but it does not need any manual work to help the tracking and does not search for the closest frame in the source sequence, which enables it to run in real-time. In addition, Dale et al. [2011] do not render novel faces but re-time the source video using dynamic time warping and blend the source and the target images directly. Therefore, their method also highly relies on the similarity between the source video and the target video. Our method builds a 3D face model from the source image at initialization and then renders it into the target. It maximally reduces the dependence on source input. Recently, *learning-based* methods were proposed to use CNN [Korshunova et al. 2017] or autoencoder [Deepfakes 2019] to learn face representations under various poses, expressions, and lighting conditions. If enough training data can be collected, these methods can produce robust and realistic results with proper post-processing. However, collecting sufficient, often large-scale, training data for specific faces is non-trivial and time-consuming, or even infeasible for some cases (e.g., legacy face videos). Furthermore, the face images they produce are generally low resolution, while our method does not have the above issues. Recently, [Nirkin et al. 2018] proposed to train a generalized face segmentation network on large face datasets, so that no additional data was required for face swapping during testing. Similar to image-based methods, this method cannot guarantee the temporal smoothness of the output sequence.

3 APPROACH OVERVIEW

Our method takes a single source portrait image and a target video clip as inputs, and outputs a video-realistic video clip with the swapped source face. Our approach consists of several steps as shown in Figure 2. We briefly introduce our pipeline in this section and describe the details of each step in the following sections.

We first reconstruct the 3D face models, albedos, illuminations, and head poses from the source image and each frame of the target video (Section 4). Each face model is further decomposed into a coarse model representing facial expressions (Section 4.1) and vertex displacements representing skin wrinkles (Section 4.2). Then, we synthesize a novel source face mesh with target expression and

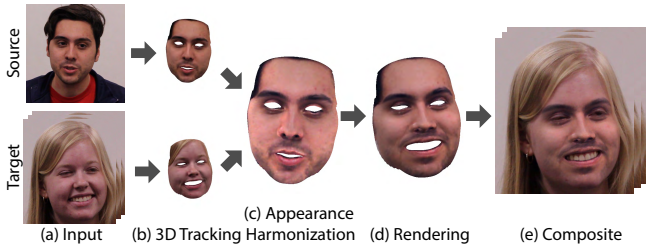


Figure 2: From the input source image and target video (a), our system captures fine-scale 3D facial performance (b). The appearance of the source face is harmonized to match the target video (c). A novel face is rendered with the source identity, harmonized appearance under the target conditions (d). The rendered face is blended into the warped target frame (e).

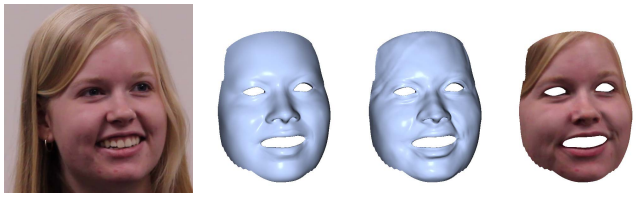


Figure 3: From an input image (a), a coarse mesh (b) is reconstructed, and augmented with vertex displacements (c). The re-rendering with the captured albedo and illumination is shown in (d).

predicted wrinkle dynamics (Section 5). We adapt the albedo of the source face to that of the target face through solving a Poisson equation (Section 6.1). The appearance is further harmonized by injecting matched noise to compensate for the different shot conditions (Section 6.2). A novel face image can be rendered by combining the novel mesh and harmonized appearance of the source face, with the illumination and head pose of the face in each target frame. Finally, we warp the target frame according to the key points of the rendered face and blend them seamlessly (Section 7).

4 FACE TRACKING

In this section, we first describe how to capture coarse-scale 3D facial performance in Section 4.1. Then, we describe how to refine the captured model via shape-from-shading to obtain fine-scale details. The same method is applied to both the source image and every frame in the target video.

4.1 Coarse Face Modeling

Reconstruction of a 3D face model from a 2D image is an intrinsically ill-posed problem. Similar to [Shi et al. 2014; Wang et al. 2016], we fit a parametric 3D face model to the image. Specifically, we use the FaceWarehouse dataset [Cao et al. 2014b] to construct a reduced bilinear core tensor C_r. A specific 3D face model can be created by multiplying the tensor C_r with a 50-dimensional identity vector

α and a 25-dimensional expression vector β. We also predefine 73 key points on the face model and run the local binary feature (LBF) based regression method to automatically track the corresponding 2D facial landmark locations from the image. Then, a coarse face model can be estimated by minimizing the distance between the detected 2D landmarks and the projected 3D key points:

$$E_{tracking} = \sum_{i=1}^{73} \|[R(v_i \times_2 \alpha \times_3 \beta) + t] - p_i\|_2^2, \quad (1)$$

where v_i and p_i are the corresponding sparse 3D and 2D key point pairs, R and t are the rotation and translation of the head, respectively. Figure 3b shows the coarse face model captured from the input in Figure 3a.

4.2 Shape from Shading Refinement

The reconstructed face model contains 5.6K vertices and 33K triangle faces presenting large-scale facial deformations. We further refine it with fine-scale details by estimating per-vertex displacements by employing the algorithm in [Ma and Deng 2019b]. Since the resolution of the mesh is too low to faithfully reproduce the subtle details of the input image, we recursively apply 4-8 subdivision [Velho and Zorin 2001] to the mesh until the vertices and pixels have an approximately one-to-one mapping. Note that the subdivision is applied to the source face model only and then copied to the target face model, so that the source and the target face share the same topology.

Inspired by the work of [Ma and Deng 2019b], we encode the fine surface bumps as the displacements along surface normals and recover them jointly with the albedo and illumination using shape-from-shading [Horn 1975]. We assume human faces are Lambertian surfaces, and parameterize the incident lighting with spherical harmonics [Basri and Jacobs 2003]. Thus, the unknown parameters can be estimated by minimizing the difference between the input face and the synthesized face:

$$E_{shading} = \sum_{i=1}^{K} \|I_i - l \cdot SH(n_i)\rho_i\|_2^2. \quad (2)$$

Here, I_i is the sampled image gradient by projecting the i-th vertex onto the image plane according to head pose, and K is the number of vertices after subdivision. l is an unknown 9-dimensional vector for the spherical harmonics coefficients of incident lighting, and SH is the second order spherical harmonics basis functions taking a unit length surface normal n_i as the input. The vertex normal n_i is calculated using the vertex positions of itself and its 1-ring neighbor vertices. We assume that the fine face model is formulated as moving each vertex of the coarse model along its normal for a distance d_i. Hence, the unknown normal n_i of the fine model is represented as a function of variable d_i. ρ_i is the unknown face albedo at vertex i, which is initialized as the average face albedo provided by the FaceWarehouse dataset [Cao et al. 2014b]. We employ block coordinate decent algorithm to alternatively solve the unknown illumination l, albedo ρ, and displacement d.

Illumination. We use the albedo and displacement from the previous frame to estimate l. We also impose a regularization term $\left\|l^t - l^{t-1}\right\|_2^2$ to penalize sudden changes between consecutive frames. The shading energy function Eq. 2 can be reduced and solved as a highly over-constrained linear system with $K + 9$ rows and 9 columns.

Albedo. Similarly, we use the estimated illumination and displacement from the previous frame to estimate ρ. To prevent the high frequency image gradients from being interpreted as albedo changes, we incorporate a Laplacian regularization term $\|L\rho - L\bar{\rho}\|_2^2$ to adapt the albedo to be as smooth as the prior average albedo $\bar{\rho}$. L is the graph Laplacian matrix with respect to the mesh. This also leads to solving a sparse linear least square problem with $2K$ rows, K columns, and $K + 2E$ non-zero entries, where E is the number of edges in the subdivided mesh. Note that the albedo is computed only at the start of the video and remains fixed thereafter.

Displacements. By substituting the estimated illumination and albedo, the shading energy function Eq. 2 is still non-linear and under-constrained in terms of displacements d_i. Here we impose two additional constraints. For a C^2 surface, its local displacements should change smoothly. Similar to albedo, a smoothness constraint is applied: $\|Ld\|_2^2$, where L is the same graph Laplacian matrix. We assume that the coarse mesh already provides a good approximation of the ground truth, thus a regularization constraint is applied: $\|d\|_2^2$. The weight for the smoothness constraint and the regularization constraint are set to 30 and 5, respectively. Figure 3c shows the refined mesh using vertex displacements, and Figure 3d shows the rendering result using the estimated albedo and illumination.

5 FACE SWAPPING

The task of face swapping is defined as replacing the face in the target video with the face from the source image while retaining the facial performance of the target actor. The hair, body and background in the target video are intact. Unlike recent deep learning based face swapping methods that learn face features in 2D images, our method also takes care of 3D face mesh swapping. We break the face geometry into large-scale expression and fine-scale wrinkles, and transfer them separately from the target to the source.

Coarse mesh swapping. The coarse mesh of the swapped face is represented as the combination of the identity of the source face and the expression of the target face. The mesh is generated by multiplying the FaceWarehouse core tensor by the identity parameter α^S of the source image and the expression parameter β^T of each frame in the target video: $\tilde{M}_t^S = C_r \times_2 \alpha^S \times_3 \beta_t^T$. Top right of Figure 4 shows the swapped coarse mesh for one frame of the target video. The whole mesh sequence is temporally smooth since α^S is constant and β^T changes smoothly in the expression PCA space.

Wrinkle prediction. The coarse mesh is further augmented with wrinkle details. The objective is to predict the most plausible person-specific wrinkle motions of the source actor under the target expression. We tackle this using the Laplacian Coating Transfer technique [Sorkine et al. 2004]. For the source face, we compute the Laplacian coordinates of the coarse mesh $\tilde{M}_0^S$ and the fine mesh M_0^S respectively for the initial source face reconstructed from image. The coating of the source mesh is defined as $\xi_0^S = L(M_0^S) - L(\tilde{M}_0^S)$, where L is the Laplacian operator. Suppose that there exists a frame

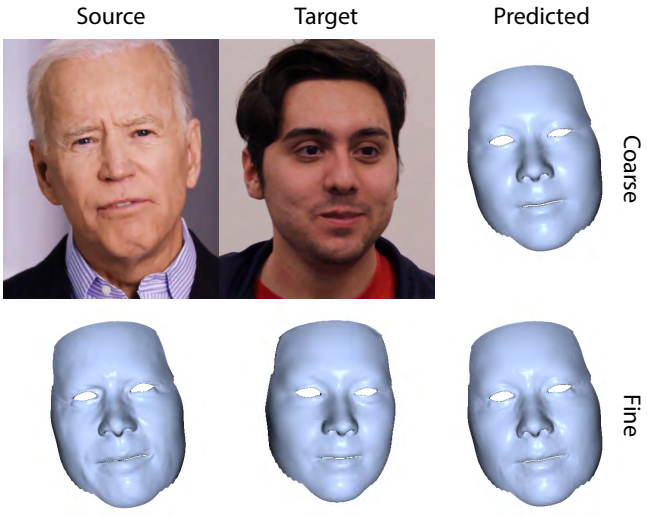


Figure 4: Face Mesh swapping from source (left) to target (middle). Top right shows the coarse mesh of the source identity performing target expression, and bottom right is the mesh with wrinkle details. Image courtesy: Joe Biden (public domain).

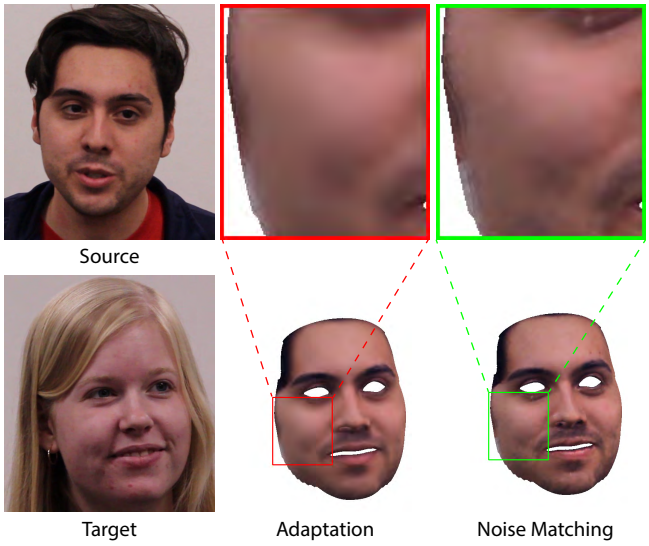


Figure 5: The appearance of the source face is harmonized to match that of the target face through albedo adaptation (middle) and noise matching (right).

in the target video that has the same expression as of the source image. Then the corresponding coating of the target can be defined similarly as $\xi_0^T = L(M_0^T) - L(\tilde{M}_0^T)$. For any frame t in the target video, the fine-scale motion $D_t^T = \xi_t^T - \xi_0^T$ is transferred to the source mesh with a local rotation R at each vertex which is the rotation of the tangent space between the coarse meshes $\tilde{M}_t^S$ and $\tilde{M}_t^T$. The coating of the source face at frame t is then predicted as $\xi_t^S = \xi_0^S + R(D_t^T)$. Finally, the fine-scale source mesh is reconstructed by solving the following inverse Laplacian:

$$M_t^S = L^{-1}(L(\tilde{M}_t^S) + \xi_t^S) = L^{-1}(L(\tilde{M}_t^S) + \xi_0^S + R(D_t^T)). \quad (3)$$

In practice, we first find a frame with the most similar expression in the target video clip to that of the source image by measuring the squared Mahalanobis distance:

$$Sim(\beta_t^T, \beta_0^S) = (\beta_t^T - \beta_0^S)^T C_{exp}^{-1}(\beta_t^T - \beta_0^S), \quad (4)$$

where C_{exp} is the covariance matrix of expression constructed from the FaceWarehouse dataset. Then the mesh from the found frame is set to M_0^T and used to compute the coating transfer. For live applications, we set the mesh from the first frame as M_0^T and keep updating it whenever a closer expression is found using Eq. 4. Bottom right of Figure 4 shows the result of coating transfer where the aging static wrinkles of the source actor are retained when performing the target expression.

6 APPEARANCE HARMONIZATION

Synthesis of a photo-realistic novel face video that combines the source face and the target background is challenging. The colors of the source face and the target face may be quite different, which leads to obvious seams along the face boundary. In addition, the source face image and the target video are usually shot under different lighting conditions. Even alpha blending or gradient domain composition may produce unrealistic results. A viable solution is to apply image harmonization to the rendered face and the target frame [Sunkavalli et al. 2010]. However, this method is not suitable for our real-time system, since it leads to the solving of very large sparse linear systems, and cannot be ported to GPU trivially. Our experiments also show that it cannot guarantee the temporal consistency of the resulting sequence. Therefore, we propose to harmonize the facial appearance in texture space. We compute the adapted albedo and the matched noise for each vertex. The values are computed only for the first several frames of the video when the albedo of the target face is also being updated simultaneously. Then, they will remain fixed during the rendering of the remaining video frames so that the rendered facial appearance can be guaranteed to be temporally consistent.

6.1 Albedo Adaptation

We compute an adapted albedo color at each vertex for face rendering. The adapted albedo is supposed to have a similar global color to the target face while having the same local gradient to the source face. This is equivalent to the solving of a Poisson equation [Pérez et al. 2003] in the texture space: the albedo values of the boundary vertices are set to the target face albedo and the albedo gradients of the inner face are set to the source face albedo gradients. The finite difference discretization of the Poisson equation yields the following discrete optimization problem:

$$E_{albedo} = \sum_{i,j \in E} \left\| (\rho_i^S - \rho_j^S) - (\hat{\rho}_i - \hat{\rho}_j) \right\|_2^2, \quad \text{s.t.} \quad \hat{\rho}_i = \rho_i^T, \quad \text{for } i \in \partial M \quad (5)$$

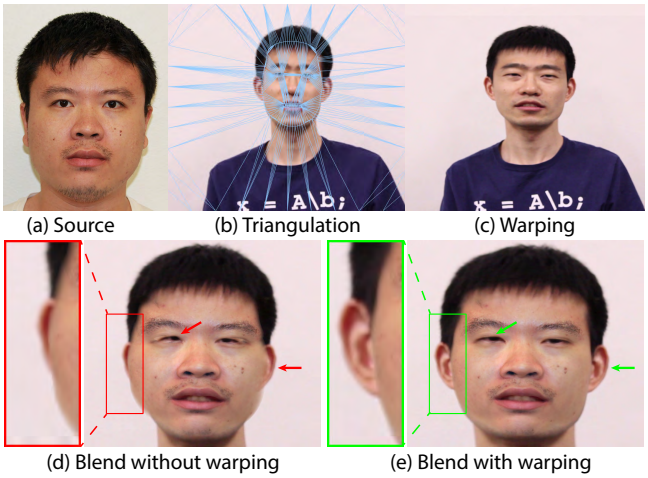


Figure 6: A target frame is triangulated using facial landmarks and boundary vertices (b). The frame is warped according to the positions of those vertices on the rendered face (c). (d) and (e) show the final blending results without and with warping.

where E denotes the edges of the face mesh, and ∂M denotes the boundary vertices of the face mesh. $\rho^S, \rho^T, and \hat{\rho}$ represent the albedo of the source face, the albedo of the target face, and the adapted albedo, respectively. Figure 5 (second column) shows the effect of the albedo adaptation.

6.2 Noise Matching

The rendered face using the adapted albedo color could be less noisy compared to the background of the target video, because of the imposed smoothing term for albedo estimation (see Section 4.2). We inject a noise color for each vertex to match the noise pattern in the target background. We first compute the noise γ in the source face and the target face, respectively, as the difference between the input image and the rendered face:

$$\gamma_i = I_i - l \cdot SH(n_i). \tag{6}$$

Then, we apply histogram matching to obtain the matched noise $\hat{\gamma}_i$:

$$\hat{\gamma}_i = \text{histmatch}(\gamma_i^S, \gamma_i^T), \tag{7}$$

where histmatch() denotes the transfer function that matches the histogram of the source noise γ_i^S with that of the target noise γ_i^T. Figure 5 (third column) shows the effect of noise matching.

7 VIDEO RENDERING AND COMPOSITION

Now we can render a novel image of the source face under the target condition and blend it to the target background. The fine-scale model M_t^S of the novel face is computed using Eq. 3. The head pose is expected to be identical to that of the target face so that the rendered face is exactly overlaid on the target face region. The final vertex position $\hat{M}$ at frame t is computed as:

$$\hat{M} = R^T M^S + t^T, \tag{8}$$

where R^T, t^T represent the rotation and translation of the target face at frame t, respectively (see Section 4).

Next, we compute the vertex normal $\hat{n}$ for the novel face model and render it with the harmonized appearance under the target illumination l^T:

$$\hat{I}_i = [l^T \cdot SH(\hat{n}_i)](\hat{\rho}_i + \hat{\gamma}_i). \tag{9}$$

A potential issue of face swapping is that the face shapes of the source and the target could greatly differ. For example, when an oblong face (Figure 6b) is swapped by a square face (Figure 6a), the eyes/ears in the background might be covered by the rendered face, which can lead to certain artifacts (Figure 6d). To alleviate the issue, we warp the background image according to the boundary and key points of the face. Specifically, we project the boundary and landmark vertices of the target face onto the image plane and use them to subdivide the image with Delaunay triangulation (Figure 6b). Since the face topology is fixed, the triangulation is applied only once and cached. Then, the same vertices of the novel face model $\hat{M}$ are projected to the target frame and used to warp the background (Figure 6c). The above operations can be efficiently done in an image quad drawing shader, where the vertex projections of the novel mesh are used as positions and those of the target mesh are used as texture coordinates. Finally, we blend the rendered face into the warped background with GPU alpha blending. Again, we pre-build an alpha map in the texture space where boundary vertices have smaller alpha values. The final composition result is shown in Figure 6e.

8 IMPLEMENTATION DETAILS

We implemented our approach in C++ and CUDA. The coarse mesh reconstruction stage runs on CPU, and the other stages run on GPU concurrently. The linear equation in illumination estimation is solved using the cuSolver. The non-linear least square problem in displacement recovery is solved by a Gauss-Newton solver in CUDA kernels. We run 10 Gauss-Newton steps for each frame, and the optimal step length is computed by a Preconditioned Conjugate Gradient (PCG) solver in 10 iterations. For albedo recovery and appearance harmonization, we solve the large sparse linear system by running the similar PCG solver for 100 iterations and match the noise histogram of 256 bins in CUDA kernels.

9 RESULTS

We demonstrate some results of our method in Figure 7 and Figure 8. In Figure 7 we swapped the same source face into multiply different target face video clips. Even though the skin color is changed to be in harmony with the target face, the face shape, eyebrows, nose and mustache of the source face remain in the result. In Figure 8 we swapped the target face by multiple different source faces. Note that face shapes and facial features such as eyebrows, nose and nevus in the results are swapped by those of the source faces, while skin color, expression and lighting are inherited from the target face. Figure 10 shows more results for faces with different genders, head poses and expressions. Please refer to the enclosed supplemental demo for video results.

All the experiments in this paper ran on a desktop computer with Intel Core i7 CPU @3.7 GHz and nVidia Geforce GTX 2080Ti GPU.

The input images and video clips were captured using a Logitech C922x Pro webcam. We shot the source images at 1080P resolution, and shot the target video at 720P resolution and 60 FPS. The coarse mesh reconstruction takes 1 ms CPU time; the mesh refinement, swapping, rendering and composition take 8 ms CPU and 18 ms GPU time. The albedo estimation and adaptation take 3 ms in total and only run at the first several frames of the video. Overall, our system ran at 55 FPS on our experimental computer.

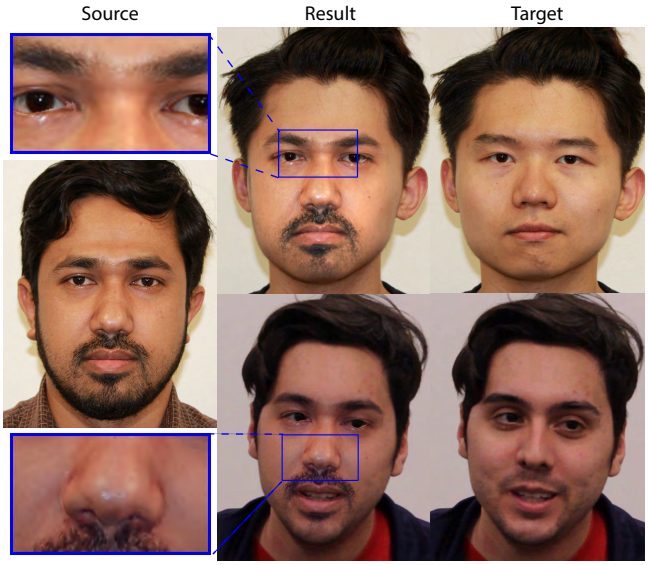


Figure 7: Face swapping results (second column) from the same source face (first column) to multiple target faces (third column). Rectangles show some examples of facial features (eye brows, nose shape, moustache, etc) are transferred from the source, while the expressions are extracted from the targets (eyebrow raising, mouth opening).

9.1 Evaluation

To quantitatively evaluate our method, we present a self swapping experiment in Figure 9. We take the first frame of the video as the source image and swap the faces in the remaining frames of the video. The heat map of the photometric error in Figure 9 visualizes the difference between the ground truth frame and the synthesized frame in RGB color space. Most regions have a good approximation with error lower than 3 in the range [0, 255]. Large errors occur at high frequency areas, such as eyebrows, or at background which are caused by background warping.

9.2 Comparisons

Previous 2D image-based methods [Bitouk et al. 2008; Kemelmacher-Shlizerman 2016] automatically select the most similar face from a large image database for face swapping. However, they cannot handle videos since different faces can be selected when the expression is changed. Thus, the results cannot be temporally consistent. Garrido et al. [2014] use videos as input for both the source and the target faces and select the most similar frame from the source video. If the source video did not contain enough variations, artifacts may occur in the results. In the following, we mainly compare our method with the state-of-the-art model-based methods and learning-based methods.

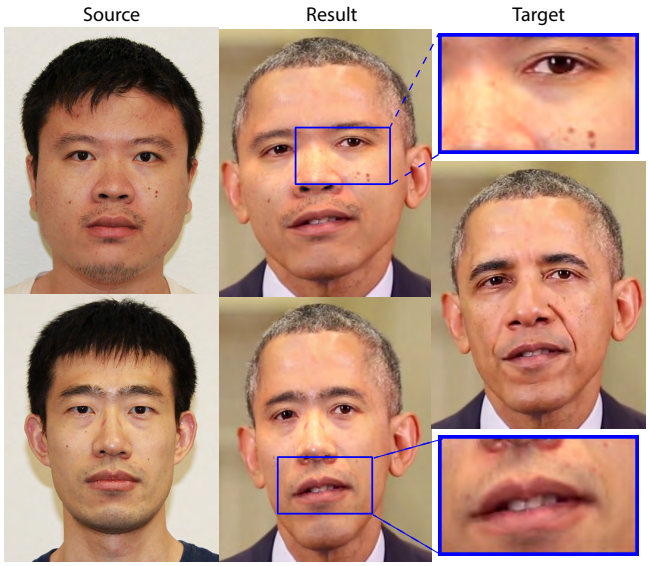


Figure 8: Face swapping results (second column) from multiple source images (first column) to the same target video (third column). Note face shapes are altered after swapping. Rectangles show some examples of facial features (lip shape, acnes, freckles, etc) are transferred in high resolution. Image courtesy: The White House (public domain).

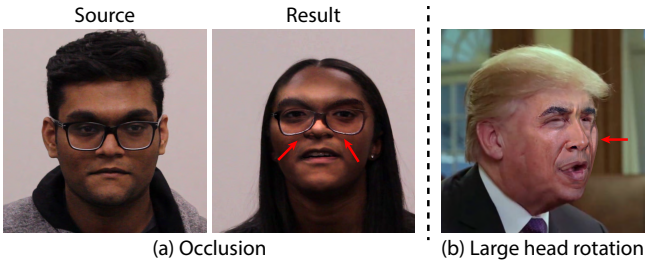
Figure 9: The photometric error of a self-swapping in which we use the first frame as the source image. Image courtesy: The White House (public domain).

Comparison with model-based methods. We compared our method with a 3D model-based method FaceSwap[2019] in Figure 10c. *FaceSwap* also fits a 3D face model for both the source and the target frames by minimizing the difference between the projected shape and the localized landmarks. However, the 3D model is only used as a proxy to warp the source face image to the target; the lighting difference between the two images are not taken into account. The rendered face may look unrealistic when the lighting conditions of the source and the target frames highly differ. This method does not take identity consistency into consideration either, therefore, the synthesized facial motion is not temporally smooth. The incoherence is better visualized in the supplemental video.

(a) Source (b) Target (c) FaceSwap (d) Deepfakes (e) Nirkin et al. (f) Ours

Figure 10: Compared to Faceswap[2019] (c), Deepfakes[2019] (d) and Nirkin et al.[2018] (e), our method (f) can change the face shape and our result contains much more facial details without the need of any training data.

Comparison with learning-based methods. We also compared our method with the state-of-the-art deep learning methods [Deepfakes 2019] (Figure 10d) and [Nirkin et al. 2018] (Figure 10e). *Deepfakes* can swap faces between two subjects. However, this method needs to train one encoder and two decoders using two large face image datasets of the two specific subjects. At runtime, the source image is passed to the encoder and then decoded by the decoder of the target subject. For comparison, we recorded two high resolution video clips (1080p) for our two subjects. We cropped the faces in a 512×512 region, yielding about 32K training images for each subject. We trained the model at batch size 64 for 100K iterations. It is obvious that the results by *Deepfakes* are much more blurring than ours, and the method of *Deepfakes* cannot change the face shape. In contrast, our method produces high resolution results with clear details and correct face shapes. Moreover, our method is more applicable and friendly to users than *Deepfakes*, since our method does not require the collecting of large-size training data of the specific faces nor expensive and time-consuming model training. [Nirkin et al. 2018] is another learning-based method that trained a deep segmentation network to guide the face swapping area. We run the code provided by the authors on the same input. As shown in Figure 10e, their result is less video-realistic compared with *Deepfakes* and our method. In the supplemental video, we show that their result sequence is also much more jumpy than the other methods.



(a) Occlusion (b) Large head rotation

Figure 11: Our method cannot effectively handle occlusions (a) or large head rotations (b).

10 DISCUSSION AND CONCLUSION

In this paper we present an automatic, real-time method to swap the facial identity and appearance in RGB videos from a single portrait image, while preserving the facial performance in terms of poses, expressions, and wrinkle motions. Our method runs fully automatic, without requiring pre-collected large-size training data of both the source and the target faces, and can create video-realistic results for faces of various skin colors, genders, ages, and expressions.

Despite our method has many advantages over previous methods, it comes at a price. Since we only use a single image to capture the source face, it is inherently ambiguous to attribute a facial feature to identity or to expression. For example, if the source image shows oblique eyebrows, two possibilities exist: the subject has oblique eyebrows in the neutral expression; or the subject has flat eyebrows but in the oblique eyebrows expression. It is very difficult to distinguish between the two possibilities even for humans. If our method reconstructs the face with oblique eyebrows in the neutral expression, the oblique eyebrows identity feature will persist for the whole rendered video. Otherwise, the oblique eyebrows expression will be replaced by the target expression such that the rendered video will not have this facial feature anymore. To overcome this limitation, a viable solution is to use a collection of images of the source face, from which a more accurate identity can be reconstructed [Roth et al. 2016].

Furthermore, our method does not swap the eyes and inner mouth of the target video. People have diverse iris, and pupils in size and color. Eyes are crucial for humans to recognize faces. With the eyes untouched, the result quality is highly affected. We would like to extend our method to swap eyes and mouth regions in future work. Facial occlusions also cannot be effectively handled by our current method. Figure 11(a) shows the glasses frame is warped after swapping due to expression change. In addition, large head poses such as side-view may produce artifacts as shown in Figure 11(b), since the facial landmark detection algorithm is not accurate for extreme poses. Artifacts may also occur when the source and the target faces have different styles of face boundaries, e.g., one with hair at forehead and one without. After solving the Poisson equation, the hair color will bleed into the inner face such that the adapted albedo seems problematic, which hampers the realism of the synthesized face. This could be tackled by employing

a face segmentation method; only the inner face without hair occlusion is swapped. In future work, we also would like to explore the possibility of hair swapping. Hair colors and styles are important features to recognize people. We believe the result quality will be highly improved if the hair can be swapped simultaneously along with the faces.

ACKNOWLEDGMENTS

This work is in part supported by NSF IIS-1524782. We would like to thank all the participants in our face swapping experiments. We also would like to thank Nirklin et al. and Rössler et al. for providing their source code and results for comparison.

REFERENCES

Hadar Averbuch-Elor, Daniel Cohen-Or, Johannes Kopf, and Michael F. Cohen. 2017. Bringing Portraits to Life. *ACM Trans. Graph.* 36, 6, Article 196 (Nov. 2017), 13 pages.

Timur Bagautdinov, Chenglei Wu, Jason Saragih, Pascal Fua, and Yaser Sheikh. 2018. Modeling facial geometry using compositional vaes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3877–3886.

Ronen Basri and David W Jacobs. 2003. Lambertian reflectance and linear subspaces. *IEEE transactions on pattern analysis and machine intelligence* 25, 2 (2003), 218–233.

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 (July 2010), 40:1–40:9.

Thabo Beeler, Derek Bradley, Henning Zimmer, and Markus Gross. 2012. Improved reconstruction of deforming surfaces by cancelling ambient occlusion. In *European Conference on Computer Vision*. Springer, Springer Berlin Heidelberg, Berlin, Heidelberg, 30–43.

Thabo Beeler, Fabian Hahn, Derek Bradley, Bernd Bickel, Paul Beardsley, Craig Gotsman, Robert W. Sumner, and Markus Gross. 2011. High-quality Passive Facial Performance Capture Using Anchor Frames. *ACM Trans. Graph.* 30, 4 (July 2011), 75:1–75:10.

Amit H. Bermano, Derek Bradley, Thabo Beeler, Fabio Zund, Derek Nowrouzezahrai, Ilya Baran, Olga Sorkine-Hornung, Hanspeter Pfister, Robert W. Sumner, Bernd Bickel, and Markus Gross. 2014. Facial Performance Enhancement Using Dynamic Shape Space Analysis. *ACM Trans. Graph.* 33, 2 (April 2014), 13:1–13:12.

Dmitri Bitouk, Neeraj Kumar, Samreen Dhillon, Peter Belhumeur, and Shree K. Nayar. 2008. Face Swapping: Automatically Replacing Faces in Photographs. *ACM Trans. Graph.* 27, 3, Article 39 (Aug. 2008), 8 pages.

Volker Blanz, Kristina Scherbaum, Thomas Vetter, and Hans-Peter Seidel. 2004. Exchanging Faces in Images. *Computer Graphics Forum* 23, 3 (2004), 669–676.

Volker Blanz and Thomas Vetter. 1999. A Morphable Model for the Synthesis of 3D Faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '99)*. ACM, New York, NY, USA, 187–194.

Chen Cao, Derek Bradley, Kun Zhou, and Thabo Beeler. 2015. Real-time High-fidelity Facial Performance Capture. *ACM Trans. Graph.* 34, 4 (July 2015), 46:1–46:9.

Chen Cao, Qiming Hou, and Kun Zhou. 2014a. Displaced Dynamic Expression Regression for Real-time Facial Tracking and Animation. *ACM Trans. Graph.* 33, 4 (July 2014), 43:1–43:10.

Chen Cao, Yanlin Weng, Stephen Lin, and Kun Zhou. 2013. 3D Shape Regression for Real-time Facial Animation. *ACM Trans. Graph.* 32, 4 (July 2013), 41:1–41:10.

Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. 2014b. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics* 20, 3 (2014), 413–425.

Yen-Lin Chen, Hsiang-Tao Wu, Fuhao Shi, Xin Tong, and Jinxiang Chai. 2013. Accurate and robust 3d facial capture using a single rgbd camera. In *2013 IEEE International Conference on Computer Vision*. 3615–3622.

Kevin Dale, Kalyan Sunkavalli, Micah K. Johnson, Daniel Vlasic, Wojciech Matusik, and Hanspeter Pfister. 2011. Video Face Replacement. *ACM Trans. Graph.* 30, 6, Article 130 (Dec. 2011), 10 pages.

Deepfakes. 2019. Retrieved May 6, 2019 from https://github.com/deepfakes/faceswap

Graham Fyffe, Andrew Jones, Oleg Alexander, Ryosuke Ichikari, and Paul Debevec. 2014. Driving High-Resolution Facial Scans with Video Performance Capture. *ACM Trans. Graph.* 34, 1 (Dec. 2014), 8:1–8:14.

Pablo Garrido, Levi Valgaert, Chenglei Wu, and Christian Theobalt. 2013. Reconstructing Detailed Dynamic Face Geometry from Monocular Video. *ACM Trans. Graph.* 32, 6 (Nov. 2013), 158:1–158:10.

Pablo Garrido, Levi Valgaerts, Ole Rehmsen, Thorsten Thormaehlen, Patrick Perez, and Christian Theobalt. 2014. Automatic Face Reenactment. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR '14)*. IEEE Computer Society, Washington, DC, USA, 4217–4224.

Pablo Garrido, Levi Valgaerts, Hamid Sarmadi, Ingmar Steiner, Kiran Varanasi, Patrick Pérez, and Christian Theobalt. 2015. Vdub: Modifying face video of actors for plausible visual alignment to a dubbed audio track. *Computer Graphics Forum* 34, 2 (2015), 193–204.

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 (May 2016), 28:1–28:15.

Paulo Gotardo, Jérémy Riviere, Derek Bradley, Abhijeet Ghosh, and Thabo Beeler. 2018. Practical Dynamic Facial Appearance Modeling and Acquisition. *ACM Trans. Graph.* 37, 6, Article 232 (Dec. 2018), 13 pages.

Yudong Guo, Juyong Zhang, Jianfei Cai, Boyi Jiang, and Jianmin Zheng. 2018. CNN-based Real-time Dense Face Reconstruction with Inverse-rendered Photo-realistic Face Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2018).

Berthold KP Horn. 1975. Obtaining shape from shading information. *The psychology of computer vision* (1975), 115–155.

Pei-Lun Hsieh, Chongyang Ma, Jihun Yu, and Hao Li. 2015. Unconstrained realtime facial performance capture. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1675–1683.

Alexandru Eugen Ichim, Sofien Bouaziz, and Mark Pauly. 2015. Dynamic 3D Avatar Creation from Hand-held Video Input. *ACM Trans. Graph.* 34, 4 (July 2015), 45:1–45:14.

Ira Kemelmacher-Shlizerman. 2016. Transfiguring Portraits. *ACM Trans. Graph.* 35, 4, Article 94 (July 2016), 8 pages.

Iryna Korshunova, Wenzhe Shi, Joni Dambre, and Lucas Theis. 2017. Fast face-swap using convolutional neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*. 3677–3685.

Marek Kowalski. 2019. FaceSwap. Retrieved December 1, 2019 from https://github.com/MarekKowalski/FaceSwap

Hao Li, Jihun Yu, Yuting Ye, and Chris Bregler. 2013. Realtime Facial Animation with On-the-fly Correctives. *ACM Trans. Graph.* 32, 4 (July 2013), 42:1–42:10.

Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh. 2018. Deep Appearance Models for Face Rendering. *ACM Trans. Graph.* 37, 4, Article 68 (July 2018), 13 pages.

Luming Ma and Zhigang Deng. 2019a. Real-Time Facial Expression Transformation for Monocular RGB Video. *Computer Graphics Forum* 38, 1 (2019), 470–481.

Luming Ma and Zhigang Deng. 2019b. Real-time Hierarchical Facial Performance Capture. In *Proceeding of ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games (I3D)*. ACM, Montreal, QC, Canada, 10 pages.

Wan-Chun Ma, Andrew Jones, Jen-Yuan Chiang, Tim Hawkins, Sune Frederiksen, Pieter Peers, Marko Vukovic, Ming Ouhyoung, and Paul Debevec. 2008. Facial Performance Synthesis Using Deformation-driven Polynomial Displacement Maps. *ACM Trans. Graph.* 27, 5 (Dec. 2008), 121:1–121:10.

Yuval Nirkin, Iacopo Masi, Anh Tuan Tran, Tal Hassner, and Gérard Medioni. 2018. On Face Segmentation, Face Swapping, and Face Perception. In *IEEE Conference on Automatic Face and Gesture Recognition*.

Patrick Pérez, Michel Gangnet, and Andrew Blake. 2003. Poisson Image Editing. *ACM Trans. Graph.* 22, 3 (July 2003), 313–318.

J. Roth, Y. Tong, and X. Liu. 2016. Adaptive 3D Face Reconstruction from Unconstrained Photo Collections. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4197–4206.

M. Sela, E. Richardson, and R. Kimmel. 2017. Unrestricted Facial Geometry Reconstruction Using Image-to-Image Translation. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 1585–1594.

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 (Nov. 2014), 222:1–222:13.

O. Sorkine, D. Cohen-Or, Y. Lipman, M. Alexa, C. Rössl, and H.-P. Seidel. 2004. Laplacian Surface Editing. In *Proceedings of the 2004 Eurographics/ACM SIGGRAPH Symposium on Geometry Processing (SGP '04)*. ACM, New York, NY, USA, 175–184. https://doi.org/10.1145/1057432.1057456

Kalyan Sunkavalli, Micah K. Johnson, Wojciech Matusik, and Hanspeter Pfister. 2010. Multi-scale Image Harmonization. *ACM Trans. Graph.* 29, 4, Article 125 (July 2010), 10 pages.

Supasorn Suwajanakorn, Ira Kemelmacher-Shlizerman, and Steven M. Seitz. 2014. Total Moving Face Reconstruction. In *Computer Vision – ECCV 2014*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Springer International Publishing, Cham, 796–812.

Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. 2015. What makes tom hanks look like tom hanks. In *2015 IEEE International Conference on Computer Vision (ICCV)*. 3952–3960.

Ayush Tewari, Michael Zollhöfer, Pablo Garrido, Florian Bernard, Hyeongwoo Kim, Patrick Pérez, and Christian Theobalt. 2018. Self-supervised Multi-level Face Model Learning for Monocular Reconstruction at over 250 Hz. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ayush Tewari, Michael Zollöfer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Theobalt Christian. 2017. MoFA: Model-based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction. In *The IEEE International Conference on Computer Vision (ICCV)*.

Justus Thies, Michael Zollhöfer, Matthias Nießner, Levi Valgaerts, Marc Stamminger, and Christian Theobalt. 2015. Real-time Expression Transfer for Facial Reenactment. *ACM Trans. Graph.* 34, 6 (Oct. 2015), 183:1–183:14.

Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. 2016. Face2face: Real-time face capture and reenactment of rgb videos. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2387–2395.

Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Niessner. 2018a. FaceVR: Real-Time Gaze-Aware Facial Reenactment in Virtual Reality. *ACM Trans. Graph.* 37, 2, Article 25 (June 2018), 15 pages.

Justus Thies, Michael Zollhöfer, Christian Theobalt, Marc Stamminger, and Matthias Niessner. 2018b. Headon: Real-time Reenactment of Human Portrait Videos. *ACM Trans. Graph.* 37, 4, Article 164 (July 2018), 13 pages.

Levi Valgaerts, Chenglei Wu, Andrés Bruhn, Hans-Peter Seidel, and Christian Theobalt. 2012. Lightweight Binocular Facial Performance Capture Under Uncontrolled Lighting. *ACM Trans. Graph.* 31, 6 (Nov. 2012), 187:1–187:11.

Luiz Velho and Denis Zorin. 2001. 4-8 Subdivision. *Comput. Aided Geom. Des.* 18, 5 (June 2001), 397–427.

Daniel Vlasic, Matthew Brand, Hanspeter Pfister, and Jovan Popović. 2005. Face Transfer with Multilinear Models. *ACM Trans. Graph.* 24, 3 (July 2005), 426–433.

Congyi Wang, Fuhao Shi, Shihong Xia, and Jinxiang Chai. 2016. Realtime 3D Eye Gaze Animation Using a Single RGB Camera. *ACM Trans. Graph.* 35, 4 (July 2016), 118:1–118:14.

Thibaut Weise, Sofien Bouaziz, Hao Li, and Mark Pauly. 2011. Realtime Performance-based Facial Animation. *ACM Trans. Graph.* 30, 4 (July 2011), 77:1–77:10.

Chenglei Wu, Takaaki Shiratori, and Yaser Sheikh. 2018. Deep Incremental Learning for Efficient High-fidelity Face Tracking. *ACM Trans. Graph.* 37, 6, Article 234 (Dec. 2018), 12 pages.

Feng Xu, Jinxiang Chai, Yilong Liu, and Xin Tong. 2014. Controllable High-fidelity Facial Performance Transfer. *ACM Trans. Graph.* 33, 4, Article 42 (July 2014), 11 pages.

Li Zhang, Noah Snavely, Brian Curless, and Steven M. Seitz. 2004. Spacetime Faces: High Resolution Capture for Modeling and Animation. *ACM Trans. Graph.* 23, 3 (Aug. 2004), 548–558.

Michael Zollhöfer, Matthias Nießner, Shahram Izadi, Christoph Rehmann, Christopher Zach, Matthew Fisher, Chenglei Wu, Andrew Fitzgibbon, Charles Loop, Christian Theobalt, and Marc Stamminger. 2014. Real-time Non-rigid Reconstruction Using an RGB-D Camera. *ACM Trans. Graph.* 33, 4 (July 2014), 156:1–156:12.

Michael Zollhöfer, Justus Thies, Pablo Garrido, Derek Bradley, Thabo Beeler, Patrick Pérez, Marc Stamminger, Matthias Nießner, and Christian Theobalt. 2018. State of the Art on Monocular 3D Face Reconstruction, Tracking, and Applications. *Computer Graphics Forum* 37, 2 (2018), 523–550.