High-Fidelity Facial Reflectance and Geometry Inference From an Unconstrained Image

SHUGO YAMAGUCHI*, Waseda University, USC Institute for Creative Technologies
SHUNSUKE SAITO*, Pinscreen, University of Southern California, USC Institute for Creative Technologies
KOKI NAGANO, Pinscreen
YAJIE ZHAO, USC Institute for Creative Technologies
WEIKAI CHEN, USC Institute for Creative Technologies
KYLE OLSZEWSKI, Pinscreen, University of Southern California, USC Institute for Creative Technologies
SHIGEO MORISHIMA, Waseda University
HAO LI, Pinscreen, University of Southern California, USC Institute for Creative Technologies

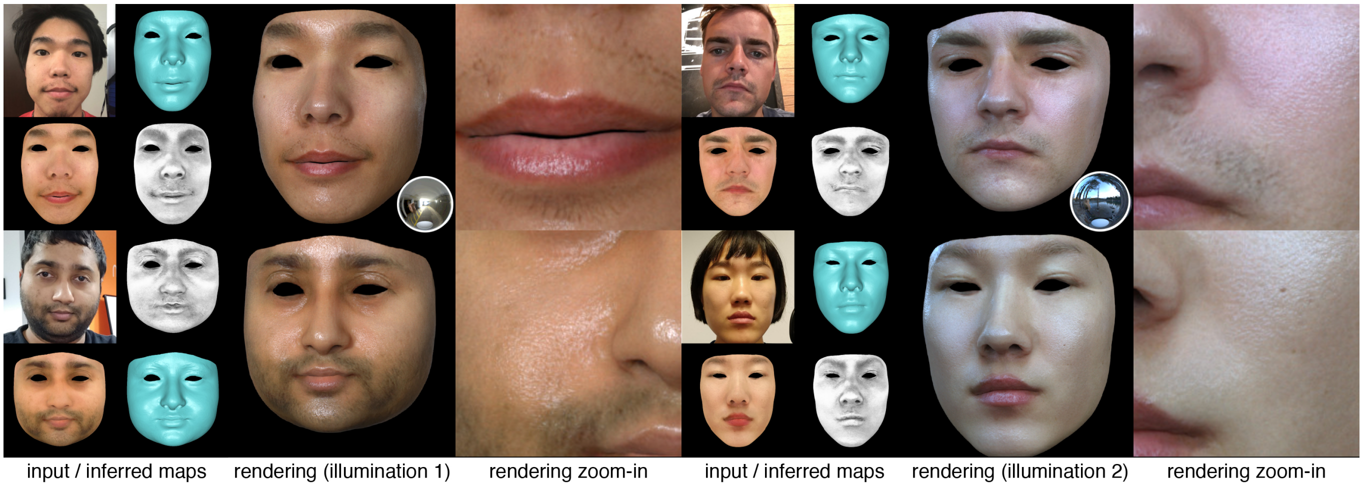

Fig. 1. Our system infers high-fidelity facial reflectance and geometry maps from a single image (diffuse albedo, specular albedo, as well as medium- and high-frequency displacements). These maps can be used for high-fidelity rendering under novel illumination conditions.

We present a deep learning-based technique to infer high-quality facial reflectance and geometry given a single unconstrained image of the subject, which may contain partial occlusions and arbitrary illumination conditions. The reconstructed high-resolution textures, which are generated in only a few seconds, include high-resolution skin surface reflectance maps, representing both the diffuse and specular albedo, and medium- and high-frequency displacement maps, thereby allowing us to render compelling digital avatars under novel lighting conditions. To extract this data, we train our deep neural networks with a high-quality skin reflectance and geometry database created with a state-of-the-art multi-view photometric stereo system using polarized gradient illumination. Given the raw facial texture map extracted from the input image, our neural networks synthesize complete reflectance and displacement maps, as well as complete missing regions caused by occlusions. The completed textures exhibit consistent quality throughout the face due to our network architecture, which propagates texture features from the visible region, resulting in high-fidelity details that are consistent with those seen in visible regions. We describe how this highly underconstrained problem is made tractable by dividing the full inference into smaller tasks, which are addressed by dedicated neural networks. We demonstrate the effectiveness of our network design with robust texture completion from images of faces that are largely occluded. With the inferred reflectance and geometry data, we demonstrate the rendering of high-fidelity 3D avatars from a variety of subjects captured under different lighting conditions. In addition, we perform evaluations demonstrating that our method can infer plausible facial reflectance and geometric details comparable to those obtained from high-end capture devices, and outperform alternative approaches that require only a single unconstrained input image.

CCS Concepts: • **Computing methodologies** → *Mesh geometry models*;

ACM Reference format:
Shugo Yamaguchi, Shunsuke Saito, Koki Nagano, Yajie Zhao, Weikai Chen, Kyle Olszewski, Shigeo Morishima, and Hao Li. 2018. High-Fidelity Facial Reflectance and Geometry Inference From an Unconstrained Image. *ACM Trans. Graph.* 37, 4, Article 162 (July 2018), 14 pages.
https://doi.org/10.1145/3130800.3130817

* indicates equal contribution

0730-0301/2018/7-ART162 $15.00
https://doi.org/10.1145/3130800.3130817

1 INTRODUCTION

Realistic digital faces are increasingly important in digital media. The capabilities of modern graphics hardware are perpetually reaching new heights, allowing for the use of effects comparable to those created using offline, state-of-the-art cinematic special effects systems in real-time, consumer-grade applications and video games. Meanwhile, the recent surge in augmented and virtual reality (AR/VR) platforms has created an even stronger demand for high-quality content for virtual environments, with applications ranging from entertainment to professional concerns, such as telepresence [Li et al. 2015; Olszewski et al. 2016; Thies et al. 2016b]. However, as immersive virtual experiences are driven by compelling human interaction, the ability to create, animate and render realistic faces plays a crucial role in achieving engaging face-to-face communication between digital avatars in simulated environments.

To render a face that appears realistic in an arbitrary virtual environment, high-quality geometry and reflectance data are required. However, acquiring this data from a real person is currently a time-consuming and cumbersome process, requiring substantial manual effort, extensive computation, and specialized capture systems operating in constrained and controlled conditions. While it would ideally be possible for a novice user to accurately model a subject's facial shape and reflectance from a single photograph (e.g. ubiquitous mobile "selfie" images), in practice, significant compromises are made to balance the amount of input data to be captured, the amount of computation required, and the quality of the final output.

We seek to efficiently create accurate, high-fidelity 3D avatars from a single input image, captured in an unconstrained environment. These avatars must be close in quality to those created by professional capture systems, with the appropriate mesoscopic geometry and reflectance attributes, yet require minimal computation and no special expertise on the part of the photographer. These requirements pose several significant technical challenges. A single photograph only provides partial data and may be taken under challenging illumination conditions. Most importantly, skin reflectance is highly complex, and as such the separation of the surface and subsurface components of the skin has only been achieved in constrained environments. Furthermore, the acquisition of accurate mesoscopic surface geometry as represented in displacement maps requires sophisticated capture hardware such as photometric multi-view stereo systems.

Less intrusive methods are based on simplifying assumptions such as the Lambertian reflectance of skin, and often make use of linear appearance models that can recover low frequency facial appearances such as the coarse shape and diffuse albedo, but fail for complex lighting conditions and detailed fine-scale facial textures, such as those containing facial hair, wrinkles, pores, and moles.

Some state-of-the-art techniques infer texture details using a database of high-resolution face textures and synthesize using a patch-based [Mohammed et al. 2009] or a neural synthesis approach [Saito et al. 2017]. However, these approaches have only been demonstrated on the reflectance aspect of the facial appearance, and thus do not provide the corresponding fine-scale geometric details needed to produce a realistic 3D rendering of the face in different views and illumination conditions. Furthermore, while [Saito et al. 2017] creates a globally consistent diffuse reflectance map from a partially occluded input texture, it replaces existing high-resolution details in the visible region, rather than preserving them and only synthesizing consistent details in the missing regions. In addition, it requires an expensive iterative optimization process, resulting in several minutes of computation time to produce the final output.

We propose a deep-learning based approach for inferring a high-fidelity set of reflectance and geometric data (including a diffuse albedo map, a specular albedo map, and medium- and high-frequency displacement maps representing mesoscopic surface details) from a single unconstrained RGB input image. To achieve robust and accurate inference in the wild, we train our model with high-resolution facial scans obtained using a state-of-the-art multi-view photometric facial scanning system [Ghosh et al. 2011]. Given the unconstrained 2D input image, which can be captured under arbitrary illumination and contain partial occlusions of the face, our process infers these high-resolution and high-fidelity geometric and reflectance maps, which can then be used to render a compelling and realistic 3D avatar in novel lighting environments, in only seconds.

Since this task is highly ill-posed, we decompose it into several more tractable problems, which are addressed by separate convolutional neural networks. In the first stage, after obtaining the coarse geometry by fitting a 3D template model to the input image and extracting an initial facial albedo map from this model, we use networks that estimate illumination-invariant specular and diffuse albedo and displacement maps from this texture. We train these networks to perform this task on arbitrary RGB images using the aforementioned 3D scans as ground-truth. While we use an architecture similar to [Isola et al. 2016], we found that several modifications to the architecture and training process were essential to enable the networks to perform this task reliably and robustly. Furthermore, we perform data augmentation using synthetic illumination conditions and simulated diffuse texture variations obtained from a pre-existing facial photograph database to make the networks more robust to appearance variation in the input images.

In the second stage, the inferred maps, which may have large missing regions due to occlusions in the input image, are passed through networks trained to perform texture completion, synthesizing full diffuse and specular albedo and displacement maps with globally consistent skin features. While a naive approach to this problem would make use of the natural symmetry in the human face to complete missing regions, human faces are not perfectly symmetric, as they contain fine-scale features (e.g., moles, hairs) that are not seen on the opposing side. We demonstrate that globally coherent high-fidelity textures can be obtained using a multi-resolution image-to-image translation network, in which latent convolutional features are flipped so as to achieve a natural degree of symmetry while maintaining local variations. In the third stage, a network is used to refine the results of the texture completion process to infer additional details in the completed regions.

Finally, we employ a convolutional neural network that performs super-resolution on each of the computed 512×512 pixel resolution textures to increase their overall resolution to 2048×2048, thereby further augmenting the level of detail in the final maps.

Using our approach, it is now possible to robustly infer realistic and accurate high-fidelity mesoscopic-level facial reflectance and geometric details from unconstrained images containing significant occlusions and arbitrary illumination conditions. The resulting data can be used with the fitted 3D model to render high-fidelity avatars in different lighting conditions and from arbitrary viewpoints. Furthermore, the resulting avatars have specific features such as facial hair, moles, and other fine-scale facial details unique to the captured subject. Once trained, our models can produce this data in only seconds, with quality comparable to that obtained from much slower and more cumbersome active-illumination capture systems. We thus present the following contributions:

- A system for obtaining a complete set of geometric and reflectance maps from a single input image. We demonstrate that the proposed technique outperforms the state-of-the-art in terms of robustness under challenging conditions, appearance preservation, and the ability to handle large appearance variations (such as facial hair or specific fine-scale features).
- The demonstration and evaluation of how our approach makes this highly ill-posed problem tractable, by performing the initial inference and texture completion using separate networks, each trained on high-fidelity 3D scans obtained using a multi-view photometric facial capture system. We describe how the architecture, training data and procedure, and data augmentation techniques are carefully chosen so as to make it possible to train these networks to robustly and accurately infer an arbitrary subject's facial appearance.
- A multi-resolution, symmetry-aware texture completion and refinement technique designed to handle the high resolution and complexity of the training data. Our approach maintains a plausible degree of symmetry in the resulting textures consistent with that seen in human faces, yet is consistent with the data observed in the visible regions.

2 RELATED WORK

2.1 Facial Reflectance and Geometry Capture.

High-Fidelity Capture. Photorealistic facial appearances can be captured by specialized hardware in controlled environments with camera arrays, e.g. the Light Stage [Debevec et al. 2000; Ghosh et al. 2011; Graham et al. 2013a; Ma et al. 2007a]. Though restricted to studio environments, such techniques have enabled production-level measurement of lighting and appearance maps, e.g. diffuse albedo, specular maps, bump maps, subsurface scattering, etc., which can be used to create realistic digital humans [Alexander et al. 2009; The Digital Human League 2015; von der Pahlen et al. 2014]. The appearance captured using such techniques can also be used with videos of the subject performing dynamic expressions to achieve high-fidelity performance capture [Fyffe et al. 2014]. Haro et al. [2001] synthesize the full-face skin structures from partial data with a high degree of accuracy. Cao et al. [2015] perform local regression of medium-scale details (e.g. dynamic wrinkles caused by facial expressions) using captured high-resolution geometry as training data. Once trained, their method scales well to new users without additional training. Optical acquisition devices and elastomeric sensors have also been introduced to the capturing pipeline for modeling facial microstructure details [Graham et al. 2013b; Johnson et al. 2011] and skin microstructure deformations [Nagano et al. 2015]. Beeler et al. [2010; 2011] applied shape from shading to emboss high-frequency skin shading as hallucinated mesoscopic geometric details for skin pores and creases. In dynamic face capture, fine-scale facial appearance can be recovered using photometric stereo techniques, e.g. photometric scene flow [Gotardo et al. 2015], spherical gradient illumination [Wilson et al. 2010] and polynomial displacement maps [Ma et al. 2008]. However, such systems require multiple images from a stereo capture, meaning that they cannot be applied to legacy content such as unconstrained images and online videos.

Linear Modeling. Modeling facial appearance variations as a linear combination of multiple bases has proven to be a popular and effective method for representing faces. Turk and Pentland [1991] present Eigenfaces for face recognition, which is one of the earliest works to represent facial appearance using a linear model. The active appearance model (AAM) proposed by Edwards et al. [1998] is another widely-adopted framework that employs a similar concept, in which faces are represented as a linear combination of both shape and appearance. It has inspired several important works in the domains of image alignment [Matthews and Baker 2004; Romdhani and Vetter 2005] and appearance retrieval [Donner et al. 2006]. The seminal work by Blanz and Vetter [Blanz and Vetter 1999] put forward the concept of a morphable model for representing 3D textured faces. By leveraging Principal Component Analysis (PCA), they first transform the shape and texture of example faces into a vector representation and estimate the coefficients of a linear basis for fitting the model to the input image. This approach is useful not only for appearance and expression modeling, but also for pose and expression normalization for face recognition [Zhu et al. 2015]. Extensions of morphable models have been developed by exploiting Internet images [Kemelmacher-Shlizerman 2013; Kemelmacher-Shlizerman and Seitz 2011] and large-scale facial scans [Booth et al. 2016]. While computationally efficient, PCA-based models are limited by the linear space spanned by the training samples, and thus are incapable of capturing fine-scale details or large variations in facial appearance.

Capturing from Unconstrained Images. Inferring local surface details using shape-from-shading is a well-established technique for unconstrained geometry capture [Barron and Malik 2015a; Glencross et al. 2008; Langer and Zucker 1994], and has been employed in digitizing human faces [Garrido et al. 2013; Kemelmacher-Shlizerman and Basri 2011; Shi et al. 2014]. However, the fidelity of the inferred details is limited by the illumination conditions of the given input images, that are often captured under unconstrained settings.

There has been a substantial effort towards the goal of making facial digitization more accessible. Monocular systems that record multiple views have been investigated to generate seamless texture maps for digital avatars [Cao et al. 2016; Ichim et al. 2015; Shi et al. 2014; Suwajanakorn et al. 2014; Thies et al. 2016a]. [Wu et al. 2016] improve the quality and robustness of monocular face capture by introducing local constraints based on the anatomy of the face so as to better capture details that are difficult to capture and express using traditional blendshape models. In the case that only a single image is available, Kemelmacher-Shilzerman and Basri [Kemelmacher-Shlizerman and Basri 2011] leverage shading information and the closest existing reference models to estimate both facial geometry and the albedo map. Barron and Malik [Barron and Malik 2015b] utilize a hybrid approach to produce a reasonable estimate of shape, surface normals, reflectance and illumination under

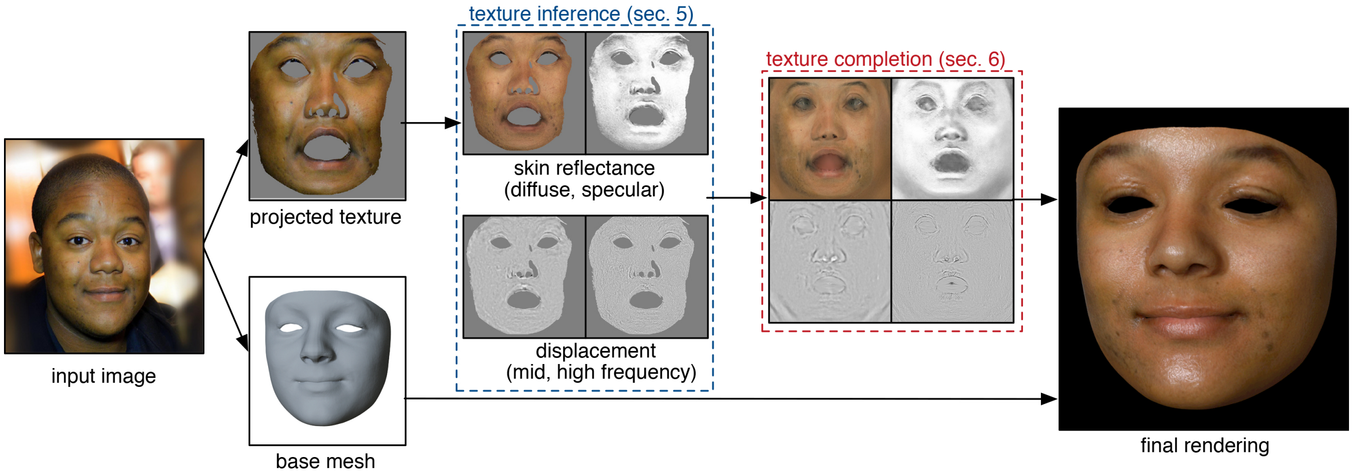

Fig. 2. System Overview. Given an unconstrained input image (left), the base mesh and corresponding facial texture map are extracted. The diffuse and specular reflectance, and the mid- and high-frequency displacement maps are inferred from the visible regions (Sec. 5). These maps are then completed, refined to include additional details inferred from the visible regions, and then upsampled using a super-resolution algorithm (Sec. 6). The resulting high-resolution reflectance and geometry maps may be used to render high-fidelity avatars (right).

a series of preset priors. [Liu et al. 2017] provide a comprehensive evaluation of the impact of several important factors, such as the number of facial landmarks and mesh vertices used, when performing cascaded regression to reconstruct 3D face shapes from a single RGB image. Li et al.[2014] take advantage of intrinsic image decomposition techniques to decouple the estimation of the specular and diffuse components of the human face. While the aforementioned techniques succeed in generating high-quality appearance models, they cannot infer the fine-scale reflectance and geometry in unseen regions. Recently, unsupervised or weakly supervised learning on facial geometry and reflectance has been proposed using color consistency [Kim et al. 2018; Sela et al. 2017; Tewari et al. 2017a,b] or synthetic data [Bradley et al. 2017; McDonagh et al. 2016; Richardson et al. 2016, 2017; Sela et al. 2017; Sengupta et al. 2017] as an additional supervisory signal.

2.2 Texture Synthesis and Image Completion

Many textures can be synthesized given a small exemplar patch using approaches based on the Markov Random Field model, as the statistical features of local regions of the texture are quite similar to all others across the entire image [Wei et al. 2009]. State-of-the-art texture synthesis techniques use various non-parametric exemplar-based techniques, such as synthesizing textures by assembling individual pixels [Efros and Leung 1999; Wei and Levoy 2000] or stitching patches [Efros and Freeman 2001; Kwatra et al. 2003; Lasram and Lefebvre 2012] of the exemplar; progressively refining the texture using a global optimization [Han et al. 2006; Kwatra et al. 2005]; or by computing high-dimensional appearance vectors for each exemplar pixel exemplar and performing synthesis in this space [Lefebvre and Hoppe 2006]. In general, however, such texture synthesis techniques only work for stochastic textures, such as micro-scale skin structures [Haro et al. 2001], and cannot be trivially applied to medium- or fine-scale facial details, as they are highly structured in addition to exhibiting local consistency. Li et al. [2007] hallucinate high frequency details from low-resolution input using a patch-based Markov network. However, the results remain blurry and missing regions cannot be inferred. Mohammed et al. [2009] generate novel faces from a random patch by combining both global and local models. Although the synthesized faces look realistic, noisy artifacts are introduced in high-resolution images. A statistical model for synthesizing detailed facial geometry has been introduced by Golovinskiy et al. [2006], but it has only been demonstrated in the geometric domain.

2.3 Deep Learning Based Image Synthesis

The advent of deep learning and its astonishing success in tasks such as image classification and face recognition has led to recent efforts to apply these networks to the task of generating images [Kulkarni et al. 2015; Radford et al. 2015]. While early efforts suffered from artifacts such as blurry images, limited resolution and little control over the synthesized image, recent efforts making use of Generative Adversarial Networks [Goodfellow et al. 2014] have led to a substantial increase in the quality of images generated using deep learning techniques, compared to networks trained using only more conventional loss metrics (such as the L1 or L2 loss on reconstructed images). In such efforts, a discriminator network is trained in conjunction with the generator, such that the discriminator learns to distinguish between real images and synthetic images created by the generator. Using loss values obtained from this discriminator results in more sophisticated criteria by which to judge the synthesized images, and thus teach the generator to synthesize higher-quality images from a distribution that more closely reflects the manifold of natural images.

However, GAN-based networks are more difficult to train, and typically fail to generate high-quality images beyond a very low resolution. While recent progress has been made [Karras et al. 2017] in synthesizing high-resolution images using an adversarial framework, it has still proven quite difficult to control precise details in the synthesized images (such as the expression in an image of a face). In recent work by Isola et al. [2016], however, GAN training has proven to improve the quality of the output and resolution for image-to-image translation tasks, in which there is a direct correspondence between pixels of the output image and those of an input image used to guide the synthesis (e.g., when synthesizing cityscape images from a semantic label map of an image). This work makes use of a conditional GAN framework, in which the discriminator is provided the per-pixel label map and must determine whether the corresponding image is real or synthesized. Olszewski et al. [2017] employs an architecture derived from this image translation framework [Isola et al. 2016] to infer dynamic facial textures from a sequence of images of a subject making a variety of facial expressions for the purpose of facial performance retargeting, but this work does not recover the individual surface and subsurface reflectance maps or the underlying mesoscopic geometry. We use an architecture similar to Isola et al. [2016] in our work, as we synthesize the resulting reflectance and geometry texture maps based a texture extracted from the input image that is in the corresponding UV space. However, substantial changes to the architecture and training process were required to achieve our desired goal.

Neural networks have also been used to infer the reflectance properties of general objects [Aittala et al. 2016]. In the context of inferring facial appearance, Duong et al. [2015; 2015] propose a nonlinear replacement of the AAM which leverages Deep Boltzmann Machines to capture both non-linearity and large variations of shape and texture. Pathak et al. [2016] introduce an encoder-decoder architecture that is conditioned on content for general image inpainting task. Iizuka et al. [2017] further incorporate both a local and a global discriminator to synthesize high-quality local details that are consistent with global background. A similar approach is used by [Li et al. 2017] for face inpainting to enhance local and global coherency. Yeh et al. [2017] iteratively search the closest embedding of a corrupted facial image in the latent space learned by a deep generative model to achieve realistic inpainting.

Recently, style transfer techniques using deep neural networks [Gatys et al. 2016, 2015] have demonstrated the capacity to combine the content of an image with a target style while preserving the structure of key visual features in the content image. Rather than synthesizing images using a forward pass through a network trained for a specified image synthesis task, these approaches iteratively modify an image passed through a pre-trained network using the feature activations of this network as guidance for the synthesis process. This ensures that a subset of these features for the modified image closely match those of a style image (such as an impressionist painting) while retaining the general content of the initial image. Inspired by the idea of defining style as mid-layer feature correlations of a neural network [Gatys et al. 2016, 2015], Saito et al. [2017] model the facial texture as a convex combination of "style" features extracted from high-resolution face database, thereby achieving photorealistic texture inference from a partial view. Hu et al. [2017] further extend this approach to generate a full-head digital avatar from a single image. Though Saito et al. [2017] have achieved photorealistic quality, their inference requires a slow and intensive iterative optimization for texture synthesis.

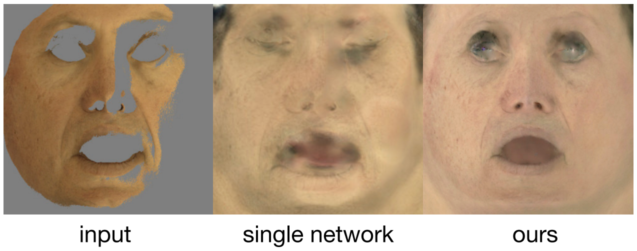

Fig. 3. Solving the described sub-tasks separately makes the complete texture inference pipeline more tractable, allowing us to generate highly plausible output. Directly generating a complete texture map from a partial input with a single network produces significantly inferior results.

Our method, on the other hand, can achieve comparable quality with [Saito et al. 2017] at a speed that is close to real time. In addition, our method is capable of inferring a much richer set of texture maps (diffuse albedo, specular albedo and displacement maps) unlike a significant body of previous techniques that are limited to diffuse albedo prediction under the assumption of Lambertian surface reflectance.

3 OVERVIEW

Our system pipeline is illustrated in Fig. 2. Given a single input image captured in unconstrained conditions, we begin by extracting the base mesh of the face and the corresponding texture map obtained by projecting the face in the input image onto this mesh. This map is passed through 2 convolutional neural networks (CNNs) that perform inference to obtain the corresponding reflectance and displacement maps (Sec. 5). The first network infers the diffuse albedo map, while the second infers the specular albedo as well as the mid- and high-frequency displacement maps. However, these maps may contain large missing regions due to occlusions in the input image. In the next stage, we perform texture completion and refinement to fill these regions with content that is consistent with that found in the visible regions (Sec. 6). Finally, we perform super-resolution to increase the pixel resolution of the completed texture from 512 × 512 into 2048 × 2048. The resulting textures contain natural and high-fidelity details that can be used with the base mesh to render high-fidelity avatars in novel lighting environments.

To obtain high-quality results, we found that it was essential to divide the inference and completion process into these smaller objectives so as to make training process more tractable, as seen in Fig. 3. Using a single network that performs both the texture completion and detail refinement on all of the desired output data (reflectance and geometry maps) produces significantly worse results than our described approach, in which the problems are decomposed into separate stages addressed by networks trained for more specific tasks, and in which the diffuse albedo is generated by a separate network than the one that generates the remaining output data.

4 TRAINING DATA

Training the networks to infer and complete the geometry and reflectance maps from the projected texture obtained from an input image requires a substantial corpus of input texture maps with corresponding ground truth reflectance and geometry maps. This data is captured with seven high-resolution DSLR cameras and a spherical LED dome, using the polarized gradient spherical illumination technique of [Ghosh et al. 2011]. The captured data includes high-resolution photographs of the subject from multiple views, sub-millimeter accurate facial geometry with a displacement map and a set of specular and diffuse albedo maps. The diffuse albedo (RGB channel) and specular albedo (single channel) respectively indicate the view-independent diffuse intensities and specular intensities with the Fresnel reflection normalized, derived from polarized spherical gradient illumination as in [Ghosh et al. 2011; Ma et al. 2007b]. Thorough definitions of these terms can be found in [Weyrich et al. 2006]. The displacement map contains the high- and medium-frequency geometric details relative to the base surface mesh, while the original high-resolution mesh is recovered by embossing the base surface with these displacements. Dense correspondences for the 3D scans are obtained with a state-of-the-art multi-view dynamic facial capture method [Fyffe et al. 2017]. These texture maps are stored in a consistent UV space such that we can learn the variation in common skin features shared by different individuals. The displacement maps are separated into medium- and high-frequency displacements. We found that this separation, which is common for facial capture [Graham et al. 2013b; Ma et al. 2008; Nagano et al. 2015], is necessary to make the training process for our networks tractable. Training using the original displacement maps, in which both the medium-frequency displacements (which contain geometric details in the range of several millimeters) and the high-frequency displacements (which may be in the sub-millimeter range) are represented in a single map leads to the high-frequency displacements being regarded as noise that is disregarded during training, and thus is not inferred properly. We separate these components using a standard Difference of Gaussians operation. The very low-frequency components of the displacement are first removed by subtracting the result of a 201×201 Gaussian filter from the raw displacements. The medium-frequency displacements are extracted by applying a 17×17 Gaussian filter to the resulting displacement maps. Subtracting these medium-frequency components from the input to this filter yields the high-frequency displacements.

Our data set includes both male and female subjects covering a variety of ages and races. The population ratio of our data is the following: $male/female = 1 : 1$, $Caucasian/Asian/African = 80 : 15 : 5$, and $Ages : 10's/20's/30's/40's/50's/60's = 5 : 40 : 25 : 20 : 5 : 5$. It consists of 329 high-resolution facial scans from 25 subjects performing up to 30 different facial expressions. We increase the data variation using several data augmentation techniques:

- Synthetic lighting augmentation: we augment the variation of the input lighting with synthetic rendering in order to obtain robust inference in the wild. To do this, we employ the ground truth facial geometry and reflectance to render the face in multiple natural HDR environments using the hybrid normal rendering [Ma et al. 2007a] and ambient occlusion technique. To simulate the natural occlusion seen in unconstrained images, we randomly perturb the head orientation and generate a visibility mask in UV space indicating which pixels are visible from this viewpoint. This visibility mask is used both at training and test time. We note that synthetic renderings have been shown to reduce the amount of training data required for and improve the robustness of learned subject-specific priors for facial expression capture [McDonagh et al. 2016]. In this work we demonstrate that similar techniques can be used to improve the quality of appearance capture results attainable using a tractable amount of high-quality ground-truth geometry and reflectance data.
- Skin diffuse albedo augmentation: we employ the Chicago Face Database (CFD) [Ma et al. 2015], which contains photographs of subjects from a wide variety of races, to increase the variety of skin tones in our dataset. We sample a number of subjects from the missing races from the CFD database and transfer the overall skin tone to the subjects in our dataset. This process is performed during training such that the distribution of skin tones in the diffuse albedo textures is similar to the skin color distribution found in the CFD. We find that this makes our approach more robust to the variety of skin tones seen in the unconstrained images used in our evaluations, particularly the darker skin tones that are underrepresented in our captured dataset.

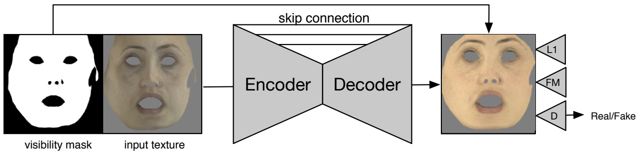

Fig. 4. Reflectance inference pipeline. The texture extracted from the input image and the corresponding visibility mask are passed through a U-net encoder-decoder framework to produce a diffuse reflectance map. Another network takes the same input and produces the specular reflectance and mid- and high-frequency displacement maps. These networks are trained using a combination of L1 and adversarial loss (D), as well as feature matching loss (FM), using features extracted from the discriminator.

5 REFLECTANCE AND GEOMETRY INFERENCE

We first adopt a pixel-wise optimization algorithm [Hu et al. 2017; Thies et al. 2016a] to obtain the base facial geometry, head orientation, and camera parameters. Using this data, we can project the face in the input image into a texture map in the UV space used in our pipeline. The non-skin region is removed in image space using a state-of-the-art semantic segmentation [Zhao et al. 2017] technique fine-tuned on the facial segmentation dataset provided by [Saito et al. 2016]. Once the input RGB texture is extracted, it may be used in the reflectance and geometry inference networks (Fig. 4) to obtain the corresponding diffuse and specular reflectance maps and the mid- and high-frequency displacement maps.

For this task, we employ a U-net architecture with skip connections similar to [Isola et al. 2016]. Such an architecture is well-suited to our task, as the skip connections between layers of the encoder and decoder modules allow for the easy preservation of the overall structure of the input image in the output image, thereby avoiding

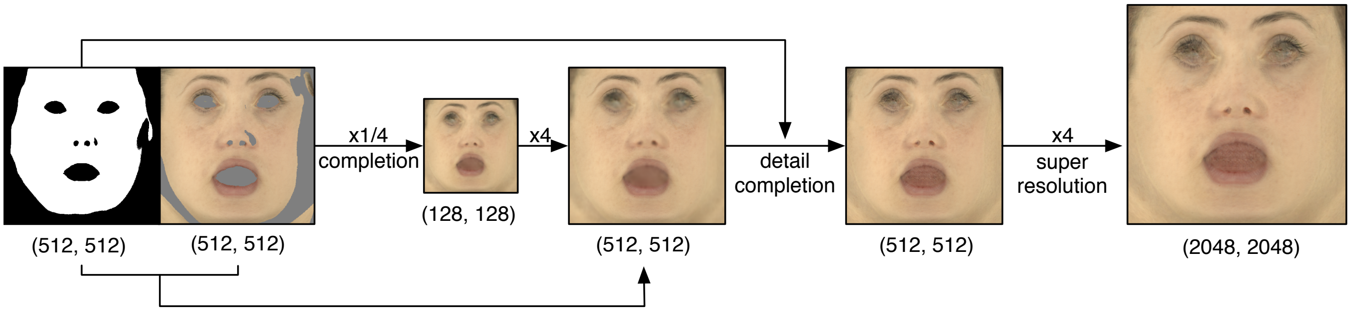

Fig. 5. Our texture completion pipeline. The inferred texture and visibility mask are downsampled by a factor of 4 and completed. The resulting low-resolution texture is upsampled to the original resolution and blended with the input texture, then passed through a network that refines the texture to add subtle yet crucial details. Finally, a super-resolution algorithm is applied to generate high-fidelity 2048 × 2048 textures.

the artifacts and limited resolutions found in more typical encoder-decoder networks. This allows the network to use more of its overall capacity to learn the appropriate transformation from the provided input to the desired output. As we perform inference in UV space, there is a direct correspondence between each pixel in the input RGB texture map and those in the inferred reflectance and geometry maps. However, we found that using this network architecture and training process is insufficient to obtain reasonable results for our task. To make this problem more tractable, we introduce several significant modifications, described below, to increase the resulting image quality and stabilize the training process:

Training Loss. During training, the L1 and GAN discriminator losses are computed only within the aforementioned visibility mask. This allows the network to focus on inferring details from only the regions that will be used in the final output. We also add a feature matching loss term using the features obtained from the discriminator, and use unconditional GAN loss, following recent efforts [Zhu et al. 2017]. We found that these modifications lead to better overall visual quality in the generated output. Note that, as we employ high fidelity rendering including ambient occlusion and subsurface scattering with hybrid normal rendering [Ma et al. 2007a] in our training data, it is non-trivial to obtain a differentiable composition on-the-fly to compute reconstruction loss, unlike [Shu et al. 2017; Tewari et al. 2017b].

Dual Networks. We use two networks with identical architectures, one operating on the diffuse albedo map (subsurface component), and the other on the tensor obtained by concatenating the specular albedo map with the mid- and high-frequency displacement maps (collectively surface components). We observed that concatenating all the data into a single input tensor leads to poor overall performance. This is because that surface and subsurface components capture different optical features of the skin, and the conflicting features interfere with one another and cause the network to fail to robustly recover each component. On the other hand, separating each component and inferring them in isolation causes instability in the training that interferes with the network's ability to recover the high-frequency displacement. We found that this separation of the diffuse component from the others results in the best overall performance, which is reasonable given that the specular reflection has a significant correlation with fine-scale details in the surface geometry (e.g., [Ghosh et al. 2011; Ma et al. 2007b] use specular analysis to recover such geometric details).

Network Architecture. To improve the accuracy of the high-frequency details, we change the stride size from 2 to 1 in the first and last convolution layers. We also add additional convolutional layers to the U-net such that the spatial dimension of the deepest layer is 1 × 1, which leads to a better encoding of the global context.

6 SYMMETRY-AWARE TEXTURE COMPLETION AND REFINEMENT

As the inferred reflectance and geometry maps often contain large missing regions due to occlusions caused by various factors (e.g., hair and non-frontal viewpoints), this inference is followed by another stage in which these missing regions are completed (Fig. 5). As with the inference stage, we find that the best results are obtained by training one network pipeline to complete the diffuse albedo and another to complete the other components (specular albedo, mid- and high-level displacement). However, we observe that completing large areas at a high resolution still does not converge to a reasonable result due to the high complexity of the learning objective. Furthermore, state-of-the-art inpainting methods work very poorly in our scenario, in which the missing region can be quite large in the case of extreme occlusions. These regions must be completed in a manner that results in natural, globally coherent facial textures free of distracting artifacts. In such cases, the convolutional layers of these networks cannot extract meaning features within their receptive fields.

Thus, we propose to stabilize the training and improve the resulting quality by dividing the inpainting problem into simpler sub-problems. The 512 × 512 resolution input textures are first resized to 128 × 128 and texture completion is performed by a network to obtain complete low-resolution textures. Second, we perform bilinear upsampling by a factor of 4 on the completed textures and blend each with the visible region in the corresponding input texture. This process is followed by detail refinement, in which these completed textures are processed to create globally coherent textures with the same level of high-fidelity details. These networks make use of the same architecture as those used for reflectance and

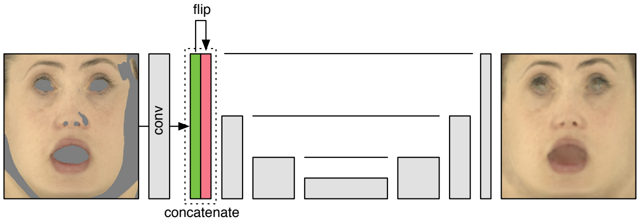

Fig. 6. Feature flipping in the latent space. The intermediate features obtained from the convolutional layers of the network are flipped across the V-axis and concatenated to the original features. This process allows the texture completion process to exploit the natural near-symmetry in human faces to infer texture maps that contain local variations but are nearly symmetric.

geometry inference (though the low-resolution completion network is modified to account for the 128 × 128 resolution input and output).

Furthermore, we leverage the spatial symmetry of UV parameterization and maximize the feature coverage by flipping intermediate features over the V-axis in UV space and concatenate them to the original features (Fig. 6). This technique allows the network to use the context provided by visible regions of the face to complete missing parts of the corresponding region on the opposite side, such as when the left half of the face is largely occluded due to a non-frontal viewpoint. We demonstrate that this feature flipping results in completed textures that do not display an uncanny degree of near-perfect symmetry, but rather contain a natural degree of symmetry as is seen in real faces. We found that this technique provided superior results to common methods for expanding the receptive field of convolution layers, such as making use of dilated convolutions or global pooling layers. Finally, the resulting 512 × 512 resolution textures are upsampled to 2048 × 2048 using a state-of-the-art super-resolution algorithm [Ledig et al. 2016].

7 IMPLEMENTATION DETAILS

We train each network using the Adam optimizer [Kingma and Ba 2014] with a learning rate set to 0.0002. In addition to the aforementioned data augmentation techniques, we perform random flipping of the input images across the V-axis to further increase the training dataset size. All training was performed on an NVIDIA GTX 1080 Ti graphics card. To train the texture completion networks, we use an occlusion mask from the input image or a random rectangular mask. We generate a mask at a random point in the image with an area ranging from $0.25 \times W \times H$ to $0.5 \times W \times H$. For the inference network, we set the weights of the L1, discriminator, and feature matching losses to 10, 1, and 0.005, respectively. For the completion network, the weights are set to 10, 1, and 0.2. For the refinement network, the weights are set to 20, 1, and 0.05.

We use three separate discriminators, one for each of the output maps, to train the network that infers the specular albedo and displacements. While this results in increased memory usage and computation when training this network, we found that superior results were obtained compared to using one discriminator that decides whether the combined output maps are real or fake. For the completion and refinement networks, we found a single discriminator operating on the entire output tensor to be sufficient.

	diffuse	specular	disp
PSNR	22.42	17.96	23.89
SSIM	0.81	0.44	0.73

Table 1. Quantitative evaluation. We measure the peak signal-to-noise ratio (PSNR) and the structural similarity (SSIM) of the inferred images for 100 test images compared to the ground truth. The inferred displacement value is computed using the output medium- and high-frequency displacement maps to recover the overall displacement.

In addition to the specified input, each network also accepts the visibility mask extracted from the initial 3D mesh fitting, as seen in Fig. 5. This allows them to better distinguish between the regions on which they must focus their capacity (such as the visible region for the initial reflectance and geometry inference, or the occluded region that must be completed for the texture completion network). We only compute and backpropagate loss for the visible region in the initial inference network, as the other regions will be completed and refined by the subsequent networks. We found that superior results were obtained from the refinement network when the adversarial and feature matching loss were backpropagated only from the occluded regions, while the L1 loss is backpropagated from the entire image. This allows the network to focus its capacity on refining the incomplete regions, which are only filled with the low-resolution output of the completion network, while maintaining the overall quality of the visible regions of the inferred reflectance and geometry maps. For the texture completion network, all losses are computed for the entire input image.

The reflectance and geometry inference networks are trained for 60,000 iterations (requiring approx. 12 hours and 6 GB of GPU memory). The texture completion networks are each trained for 60,000 iterations using the masked ground truth images, and another 30,000 iterations to fine-tune the network using the output of the initial inference networks (approx. 6 hours, 1.5 GB GPU memory). The detail refinement network is likewise trained for 60,000 iterations and fine-tuned for another 30,000 iterations using the output of the trained texture completion networks (approx. 12 hours, 4.5 GB GPU memory). The super-resolution network is trained for 1000 epochs using our training data as ground truth.

8 RESULTS

All our results are rendered with brute-force path tracing in the Solid Angle's Arnold renderer [Solid Angle 2016] with physically based specular reflection and subsurface scattering with high dynamic range image-based illumination. The resulting surface and subsurface reflectance, together with the base surface mesh and the displacement, are used to produce the final render using a layered skin reflectance model as in [The Digital Human League 2015] (see supplemental material for more details on the rendering process).

Evaluation. We quantitatively measure the ability of our system to faithfully recover the reflectance and geometry data from a set of 100 test images for which we have the corresponding ground-truth measurements. The results are seen in Table 1. We see that the system is able to recover the diffuse albedo and overall displacement quite well, though the higher complexity of the specular albedo results in a larger difference from the ground truth. However, our

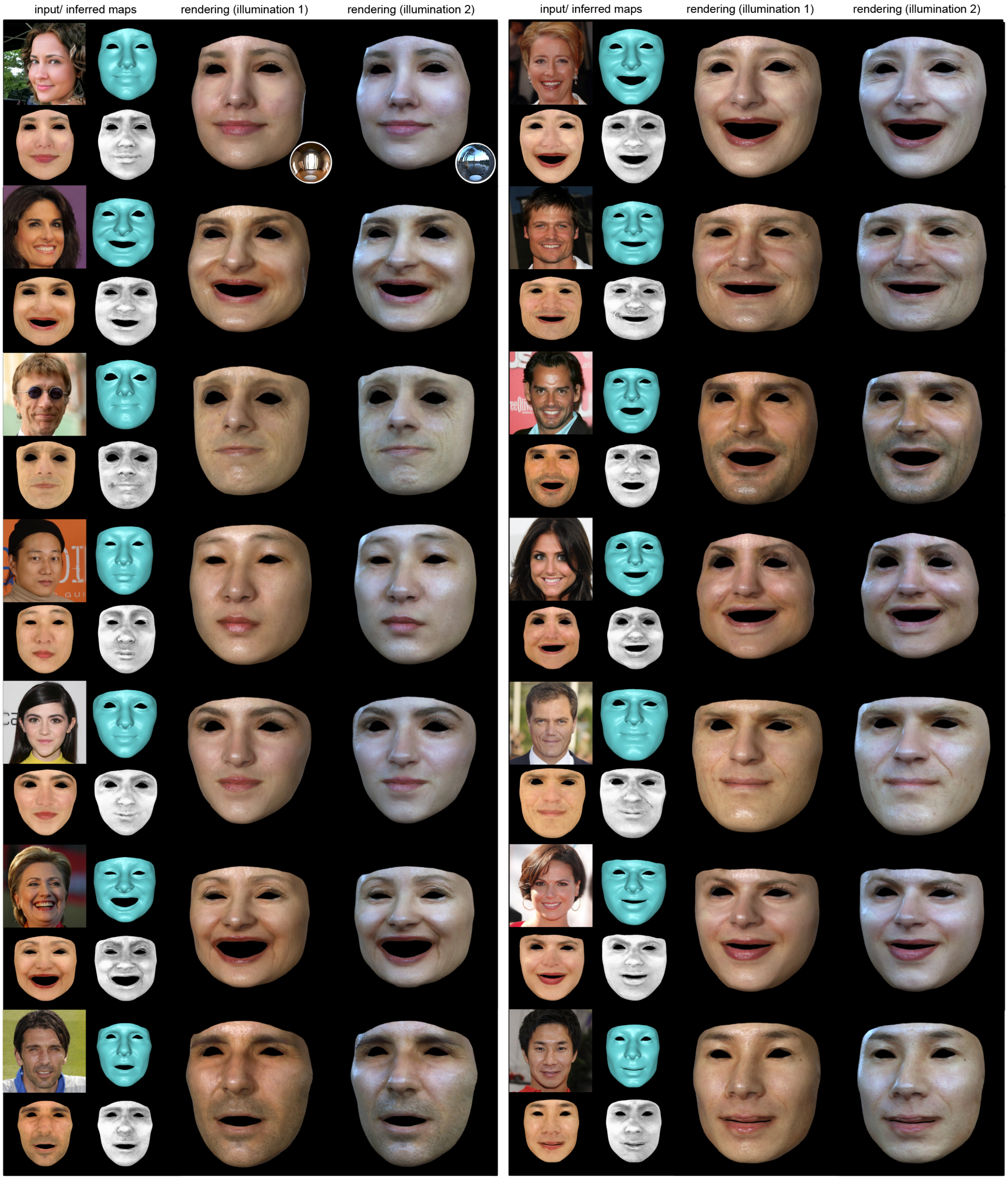
input/ inferred maps
rendering (illumination 1)
rendering (illumination 2)
input/ inferred maps
rendering (illumination 1)
rendering (illumination 2)

Fig. 7. Inference in the wild. The first column contains the input image and the corresponding inferred output applied to the base mesh. The second and third columns contain new renderings of the avatar under novel lighting conditions (the lighting environments we use are inset in the top left example renderings).

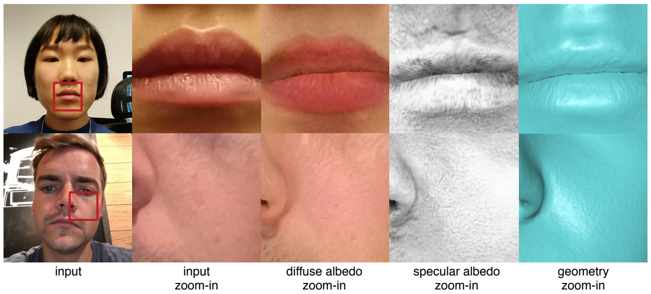

Fig. 8. Zoom-in results showing synthesized mesoscopic details.

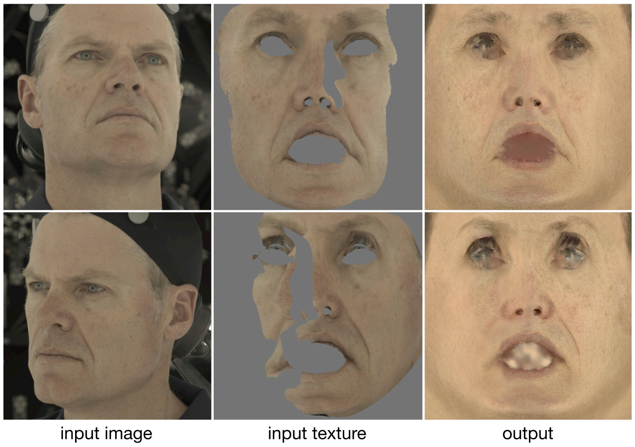

Fig. 9. Consistency of the output obtained using input images of the same subject from different viewpoints.

qualitative evaluations demonstrate that the inferred data is still sufficient for rendering compelling and high-quality avatars.

In Fig. 7, we show several results obtained using unconstrained input images from the CelebA dataset [Liu et al. 2015], with the input images, the corresponding inferred textures, and sample renderings using the inferred data. Despite the widely varying subject appearance, lighting conditions, facial expressions, and view angles with occlusions, the results demonstrate that the system is able to infer the data needed to render subjects well enough to allow for the rendering of compelling and high-fidelity avatars. In the supplemental material we provide additional single-view reconstructions with our method using a large public data set [Ma et al. 2015].

Figs. 9 and 10 demonstrate that we obtain comparable results using a single image that is captured from different viewpoints and different lighting conditions, respectively. As can be seen in Figs. 9, the missing region such as the side of the cheek is plausibly completed with our texture completion method, with appropriate natural symmetry, and is consistent with the rest of the skin. Despite the varying lighting colors, the amount of specularity, and the contrast in the images due to shadowing, the reconstructed textures display a consistent skin quality matching the subject's identity (Fig. 10). These results demonstrate that our approach can recover plausible and consistent output despite large variations in the input images, such as vastly differing viewpoints or extreme changes to the lighting environment.

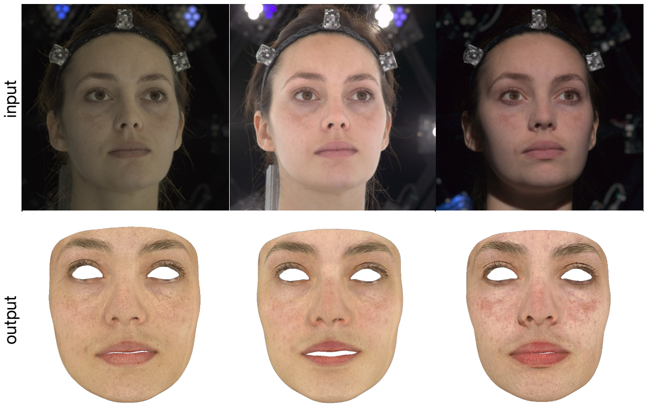

Fig. 10. Consistency of the output obtained using input images of the same subject captured under different lighting conditions.

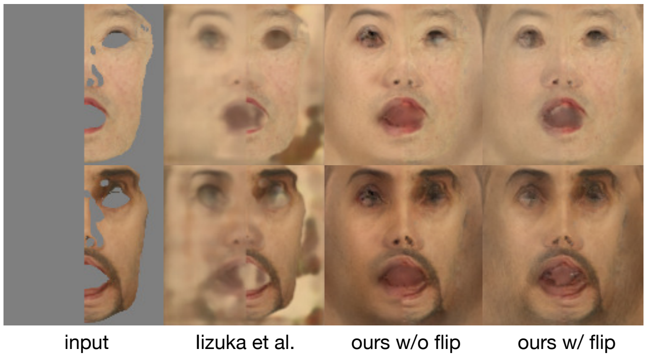

Fig. 11. Comparison with [Iizuka et al. 2017] and our network, both with and without the feature flipping layer.

We provide additional experiments in which we alter conditions such as the input view angles, lighting conditions, and expressions in the supplemental material. For more results, please watch the supplementary video.

Comparison. In Fig. 11, we compare our approach to [Iizuka et al. 2017]. Severe occlusions resulting in large missing regions in the input texture cause their method to fail to faithfully recover the entire diffuse albedo map. Our method, in contrast, is able to infer plausible and coherent data to fill the missing regions, resulting in a much more natural albedo map that is suitable for rendering a digital avatar. We provide results using our described approach both with and without the aforementioned feature flipping strategy, demonstrating the importance of this technique in producing output images that are both complete and natural.

In Fig. 12, we compare our approach to several alternatives on a variety of input subjects captured under different conditions. We show the results obtained by simply reconstructing the captured

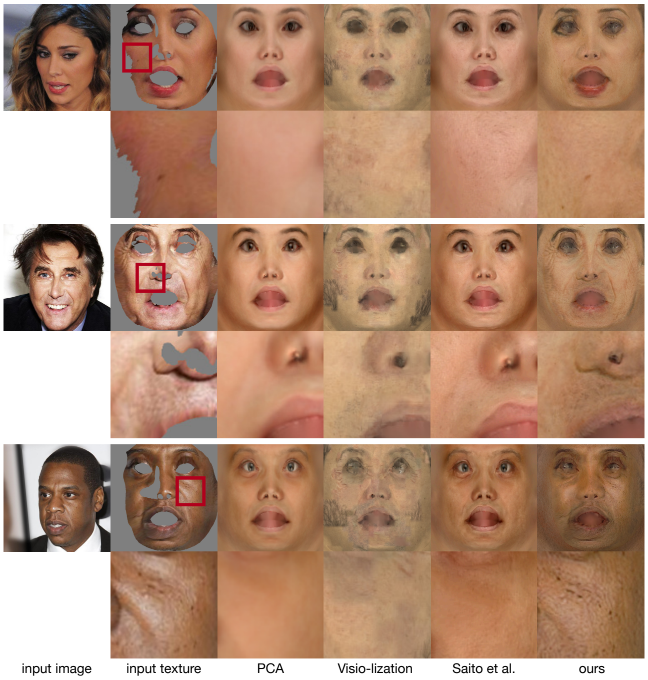

Fig. 12. Comparison with PCA, Visio-lization [Mohammed et al. 2009], and a state-of-the-art diffuse albedo inference method [Saito et al. 2017].

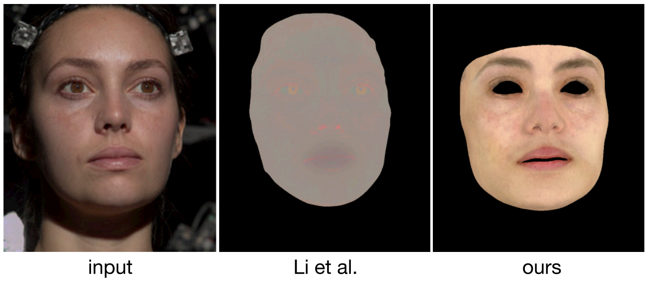

Fig. 13. Comparison of diffuse albedo inference with a data-driven intrinsic decomposition method [Li et al. 2014] (produced by original authors).

texture using the PCA coefficients obtained from the 3D face fitting process [Thies et al. 2016a] used to extract raw texture that is provided as input to our system; the results obtained using [Mohammed et al. 2009]; and the result of applying [Saito et al. 2017]. We show both the entire recovered diffuse texture as well as a close-up of a region of the texture. This clearly demonstrates our approach's ability to faithfully recover fine-scale details corresponding to the input image, resulting in more coherent and plausible facial textures than these alternative approaches. Figs. 13, 14 and 15 provide additional comparisons with the results obtained using our approach and those obtained using several recently developed facial capture techniques. As seen in the figures, our method produces significantly better skin texture (Fig. 13), sharp details (Fig. 14), and preserves distinct, person-specific details such as freckles (Fig. 15).

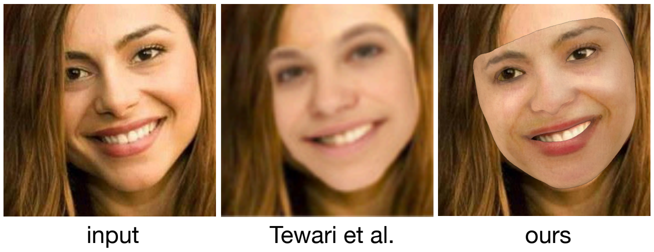

Fig. 14. Comparison of diffuse albedo inference with an unsupervised face alignment method [Tewari et al. 2017b], in which skin textures are represented by a linear basis.

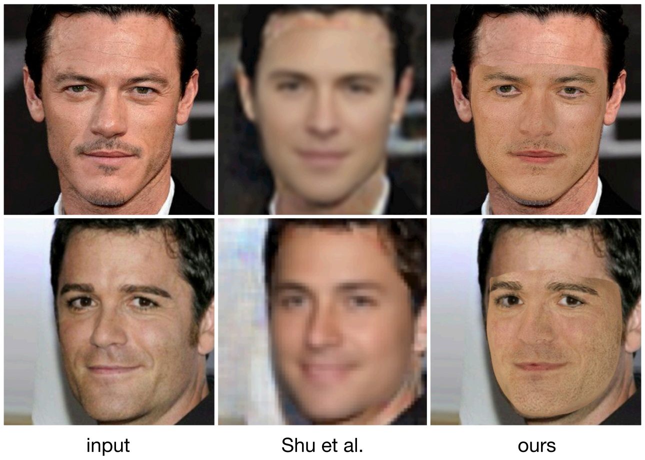

Fig. 15. Comparison of diffuse albedo inference with an unsupervised intrinsic decomposition method [Shu et al. 2017].

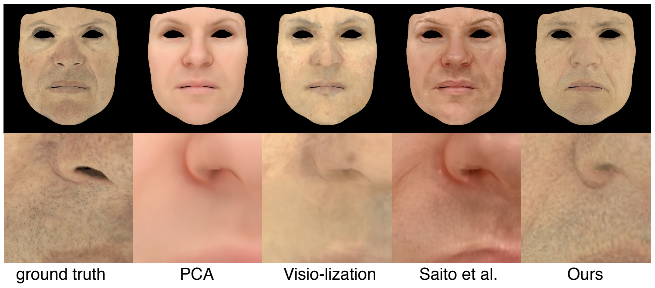

Fig. 16. Comparison with PCA, Visio-lization [Mohammed et al. 2009], and a state-of-the-art diffuse albedo inference method [Saito et al. 2017] using Light Stage ground truth data.

We also provide quantitative comparisons of the fidelity of our diffuse albedo inference with that obtained using these techniques. As seen in Table 2, our method produces albedo maps that are closer to the ground truth than any of these alternatives (see Fig. 16 for a qualitative comparison). In Fig. 17, we show rendering results using the data inferred with our approach, and compare with renderings generated using the high-fidelity data acquired directly from the

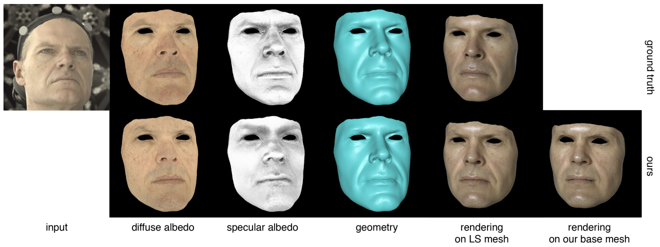

Fig. 17. Ground truth comparison using Light Stage (LS) data.

method	PSNR	RSME
[Thies et al. 2016a]	17.6354	0.1369
[Saito et al. 2017]	15.6308	0.1767
[Mohammed et al. 2009]	18.34	0.1271
ours	**19.333**	**0.1102**

Table 2. Quantitative comparison our diffuse albedo inference with several alternative methods, measured using the PSNR and the root-mean-square error (RMSE).

stage	diffuse	specular, disp
inference	8 ms	8 ms
completion	6 ms	6 ms
refinement	3 ms	3 ms
super-resolution	300 ms	300 ms

Table 3. Runtime performance for each component of our system.

multi-view stereo system used to generate our training data. The renderings with our inferred reflectance data applied to the ground truth base mesh from the Light Stage suggest that our method can capture all the reflectance data necessary to render a high-fidelity avatar. The last column shows the result using the base mesh obtained with our method. The final rendering of our single-view technique indicates comparable quality to that obtained with a Light Stage capture device.

Performance. Table 3 shows the runtime performance of each stage of our pipeline.

9 CONCLUSION

We have demonstrated the feasibility of inferring high-resolution reflectance and geometry maps using a single unconstrained image of the captured subject. Not only are these maps high-fidelity and sufficient for rendering compelling and realistic avatars, but they contain the fine details essential for preserving the likeness of the captured subject (such as pores, moles, and facial hair). This is possible in large part due to our use of high-quality ground truth 3D scans and the corresponding input images. This allows for the training of networks specially designed for the inference, texture completion and detail refinement tasks necessary to generate the data for rendering these avatars. By decomposing this problem into smaller tasks that are addressed using specific variations of the network architecture and training procedure, we are able to obtain high-resolution textures containing all the data needed to render characters with reflectance and fine-scale geometry matching the target subject. This output is comparable in quality to that obtained by [Saito et al. 2017], but is obtained in only a fraction of the time (several seconds rather than several minutes). Unlike the aforementioned approach, the output includes all the mesoscopic geometric and illumination-independent reflectance data required to produce realistic renderings under novel lighting conditions. Furthermore, our approach maintains high-resolution details in the reflectance of the input image, rather than changing the entire image to match the statistics of those in our training database, but still produces globally coherent textures. To render realistic faces, the inferred textures should not have perfect symmetry, which would result in uncanny renderings, but need to have local variations comparable to those seen in real faces. Our technique of flipping and concatenating convolutional features encoded in the latent space of our model allows us to perform texture completion in a manner that respects the natural degree of symmetry seen in the human face.

Limitations. Despite these findings, our approach has several limitations. While it is able to quickly infer high-fidelity details given a sufficiently high-resolution input image, it cannot infer these details if the input image is of very low quality or resolution, unlike the more computationally intensive transfer-based technique of [Saito et al. 2017]. Furthermore, while it can recover details such as facial stubble, which can be represented as fine details in the reflectance and geometry maps, it cannot recover other larger variations in facial appearances, such as very dense and long facial hair. Furthermore, other features that do not correspond to semantic features of the human face, such as glasses, cannot be recovered and may interfere with the fitting process used to recover the base mesh and corresponding texture map from the input image. As our ability to recover the input facial texture is limited by our ability to recover the base mesh and camera parameters using a photometric-consistency optimization, very challenging conditions in the input images, such as extreme lighting conditions or largely non-frontal viewpoints, may cause failures in this stage. In addition, strong dynamic expressions can introduce transient wrinkles that may lead to inconsistent reflectance and geometry maps for a given subject compared to those that would be obtained using an image with a more neutral facial expression. (Figure 18 contains example output produced under some of the aforementioned conditions). Very specific and unique features, such as scars, will not be recovered as accurately as when using a more cumbersome and computationally intensive approach relying on multi-view stereo capture of each new subject.

Future Work. In addition to addressing the aforementioned limitations, we believe that there are many avenues of future work in the domain of high-quality facial capture in unconstrained scenarios that could build upon our approach and make use of our high-quality facial scan database. We plan to expand our database to cover dynamic facial details, such as those caused by strong facial expressions. Extending our approach to recover dynamic fine-scale facial details from multiple input images, such as those taken in a short video or a sequence of still images, is another promising area of exploration. This would allow for the recovery of additional details

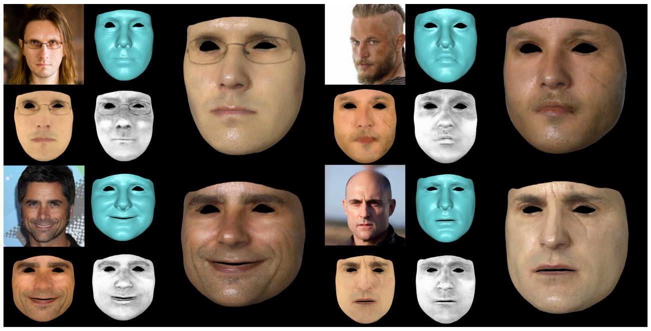

Fig. 18. Limitations. Our method produces artifacts in the presence of strong shadows (lower right) and non-skin objects due to segmentation failures (upper left). Also volumetric beards are not faithfully reconstructed (upper right). Strong dynamic wrinkles (lower left) may cause artifacts in the inferred displacement maps.

when some of the input images suffer from issues such as low resolution or extreme occlusions. It may also allow for a more accurate reconstruction of the base mesh, thereby allowing for even more accurate renderings of digital avatars using the inferred textures.

ACKNOWLEDGEMENTS

We would like to thank Bagel-mo for her grumpy face; Mr. Sirnif and Aviral Agarwal for the renderings; Carrie Sun, Jens Fursund, Mike Seymour, and Emily O'Brien for being our models; and Chen Li for the albedo inference comparisons. This research is supported by Adobe, the Andrew and Erna Viterbi Early Career Chair, the JST ACCEL Grant Number JPMJAC1602, the Waseda Research Institute for Science and Engineering, and the U.S. Army Research Laboratory (ARL) under contract W911NF-14-D-0005. The content of the information does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

REFERENCES

M. Aittala, T. Aila, and J. Lehtinen. 2016. Reflectance modeling by neural texture synthesis. *ACM Trans. Graph.* 35, 4 (2016), 65.

O. Alexander, M. Rogers, W. Lambeth, M. Chiang, and P. Debevec. 2009. The Digital Emily Project: Photoreal Facial Modeling and Animation. In *ACM SIGGRAPH 2009 Courses (SIGGRAPH '09)*. ACM, New York, NY, USA, Article 12, 15 pages.

J. T. Barron and J. Malik. 2015a. Shape, illumination, and reflectance from shading. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37, 8 (2015), 1670–1687.

J. T. Barron and J. Malik. 2015b. Shape, Illumination, and Reflectance from Shading. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2015).

T. Beeler, B. Bickel, P. Beardsley, B. Sumner, and M. Gross. 2010. High-quality single-shot capture of facial geometry. In *ACM Trans. Graph.*, Vol. 29. ACM, 40.

T. Beeler, F. Hahn, D. Bradley, B. Bickel, P. Beardsley, C. Gotsman, R. W. Sumner, and M. Gross. 2011. High-quality passive facial performance capture using anchor frames. In *ACM Trans. Graph.*, Vol. 30. ACM, 75.

V. Blanz and T. Vetter. 1999. A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques.* 187–194.

J. Booth, A. Roussos, S. Zafeiriou, A. Ponniah, and D. Dunaway. 2016. A 3d morphable model learnt from 10,000 faces. In *Proc. CVPR.* 5543–5552.

D. Bradley, T. Beeler, K. Mitchell, and others. 2017. Real-Time Multi-View Facial Capture with Synthetic Training. In *Computer Graphics Forum*, Vol. 36. Wiley Online Library, 325–336.

C. Cao, D. Bradley, K. Zhou, and T. Beeler. 2015. Real-time high-fidelity facial performance capture. *ACM Trans. Graph.* 34, 4 (2015), 46.

C. Cao, H. Wu, Y. Weng, T. Shao, and K. Zhou. 2016. Real-time facial animation with image-based dynamic avatars. *ACM Trans. Graph.* 35, 4 (2016), 126.

P. Debevec, T. Hawkins, C. Tchou, H.-P. Duiker, and W. Sarokin. 2000. Acquiring the Reflectance Field of a Human Face. In *Proc. SIGGRAPH.*

R. Donner, M. Reiter, G. Langs, P. Peloschek, and H. Bischof. 2006. Fast Active Appearance Model Search Using Canonical Correlation Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28, 10 (2006), 1690–1694.

C. N. Duong, K. Luu, K. G. Quach, and T. D. Bui. 2015. Beyond principal components: Deep boltzmann machines for face modeling. In *Proc. CVPR.* 4786–4794.

G. J. Edwards, C. J. Taylor, and T. F. Cootes. 1998. Interpreting Face Images Using Active Appearance Models. In *Proceedings of the 3rd. International Conference on Face and Gesture Recognition (FG '98).* IEEE Computer Society, 300–.

A. A. Efros and W. T. Freeman. 2001. Image Quilting for Texture Synthesis and Transfer. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '01).* ACM, 341–346.

A. A. Efros and T. K. Leung. 1999. Texture Synthesis by Non-Parametric Sampling. In *IEEE ICCV.* 1033–.

G. Fyffe, A. Jones, O. Alexander, R. Ichikari, and P. Debevec. 2014. Driving high-resolution facial scans with video performance capture. *ACM Trans. Graph.* 34, 1 (2014), 8.

G. Fyffe, K. Nagano, L. Huynh, S. Saito, J. Busch, A. Jones, H. Li, and P. Debevec. 2017. Multi-View Stereo on Consistent Face Topology. In *Computer Graphics Forum*, Vol. 36. Wiley Online Library, 295–309.

P. Garrido, L. Valgaerts, C. Wu, and C. Theobalt. 2013. Reconstructing Detailed Dynamic Face Geometry from Monocular Video. In *ACM Trans. Graph.*, Vol. 32. 158:1–158:10.

L. A. Gatys, M. Bethge, A. Hertzmann, and E. Shechtman. 2016. Preserving Color in Neural Artistic Style Transfer. *CoRR* abs/1606.05897 (2016).

L. A. Gatys, A. S. Ecker, and M. Bethge. 2015. Texture synthesis and the controlled generation of natural stimuli using convolutional neural networks. *CoRR* abs/1505.07376 (2015).

A. Ghosh, G. Fyffe, B. Tunwattanapong, J. Busch, X. Yu, and P. Debevec. 2011. Multiview Face Capture Using Polarized Spherical Gradient Illumination. *ACM Trans. Graph.* 30, 6, Article 129 (2011), 10 pages.

M. Glencross, G. J. Ward, F. Melendez, C. Jay, J. Liu, and R. Hubbold. 2008. A perceptually validated model for surface depth hallucination. *ACM Trans. Graph.* 27, 3 (2008), 59.

A. Golovinskiy, W. Matusik, H. Pfister, S. Rusinkiewicz, and T. Funkhouser. 2006. A Statistical Model for Synthesis of Detailed Facial Geometry. *ACM Trans. Graph.* 25, 3 (2006), 1025–1034.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 2672–2680.

P. F. Gotardo, T. Simon, Y. Sheikh, and I. Matthews. 2015. Photogeometric scene flow for high-detail dynamic 3d reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision.* 846–854.

P. Graham, B. Tunwattanapong, J. Busch, X. Yu, A. Jones, P. Debevec, and A. Ghosh. 2013a. Measurement-Based Synthesis of Facial Microgeometry. In *Computer Graphics Forum*, Vol. 32. Wiley Online Library, 335–344.

P. Graham, B. Tunwattanapong, J. Busch, X. Yu, A. Jones, P. Debevec, and A. Ghosh. 2013b. Measurement-based Synthesis of Facial Microgeometry. In *EUROGRAPHICS.*

J. Han, K. Zhou, L.-Y. Wei, M. Gong, H. Bao, X. Zhang, and B. Guo. 2006. Fast example-based surface texture synthesis via discrete optimization. *The Visual Computer* 22, 9-11 (2006), 918–925.

A. Haro, B. Guenterz, and I. Essay. 2001. Real-time, Photo-realistic, Physically Based Rendering of Fine Scale Human Skin Structure. In *Eurographics Workshop on Rendering*, S. J. Gortle and K. Myszkowski (Eds.).

L. Hu, S. Saito, L. Wei, K. Nagano, J. Seo, J. Fursund, I. Sadeghi, C. Sun, Y.-C. Chen, and H. Li. 2017. Avatar Digitization From a Single Image For Real-Time Rendering. *ACM Trans. Graph.* 36, 6 (2017).

A. E. Ichim, S. Bouaziz, and M. Pauly. 2015. Dynamic 3D Avatar Creation from Hand-held Video Input. *ACM Trans. Graph.* 34, 4, Article 45 (2015), 14 pages.

S. Iizuka, E. Simo-Serra, and H. Ishikawa. 2017. Globally and Locally Consistent Image Completion. *ACM Trans. Graph.* 36, 4, Article 107 (2017), 107:1–107:14 pages.

P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. 2016. Image-to-image translation with conditional adversarial networks. *arXiv preprint arXiv:1611.07004* (2016).

M. K. Johnson, F. Cole, A. Raj, and E. H. Adelson. 2011. Microgeometry Capture using an Elastomeric Sensor. *ACM Trans. Graph* 30, 4 (2011), 46:1–46:8.

T. Karras, T. Aila, S. Laine, and J. Lehtinen. 2017. Progressive Growing of GANs for Improved Quality, Stability, and Variation. *CoRR* abs/1710.10196 (2017). arXiv:1710.10196

I. Kemelmacher-Shlizerman. 2013. Internet-based Morphable Model. *IEEE ICCV* (2013).

I. Kemelmacher-Shlizerman and R. Basri. 2011. 3D face reconstruction from a single image using a single reference face shape. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33, 2 (2011), 394–405.

I. Kemelmacher-Shlizerman and S. M. Seitz. 2011. Face reconstruction in the wild. In *IEEE ICCV.* IEEE, 1746–1753.

H. Kim, M. Zollhöfer, A. Tewari, J. Thies, C. Richardt, and C. Theobalt. 2018. InverseFaceNet: Deep Monocular Inverse Face Rendering. In *Proc. CVPR.*

D. P. Kingma and J. Ba. 2014. Adam: A Method for Stochastic Optimization. *CoRR* abs/1412.6980 (2014). arXiv:1412.6980

T. D. Kulkarni, W. F. Whitney, P. Kohli, and J. Tenenbaum. 2015. Deep Convolutional Inverse Graphics Network. In *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (Eds.). Curran Associates, Inc., 2539–2547.

V. Kwatra, I. Essa, A. Bobick, and N. Kwatra. 2005. Texture optimization for example-based synthesis. *ACM Trans. Graph.* 24, 3 (2005), 795–802.

V. Kwatra, A. Schödl, I. Essa, G. Turk, and A. Bobick. 2003. Graphcut Textures: Image and Video Synthesis Using Graph Cuts. In *Proc. SIGGRAPH (SIGGRAPH '03)*. ACM, 277–286.

M. S. Langer and S. W. Zucker. 1994. Shape-from-shading on a cloudy day. *JOSA A* 11, 2 (1994), 467–478.

A. Lasram and S. Lefebvre. 2012. Parallel patch-based texture synthesis. In *Proceedings of the Fourth ACM SIGGRAPH/Eurographics conference on High-Performance Graphics*. Eurographics Association, 115–124.

C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and others. 2016. Photo-realistic single image super-resolution using a generative adversarial network. *arXiv preprint arXiv:1609.04802* (2016).

S. Lefebvre and H. Hoppe. 2006. Appearance-space texture synthesis. *ACM Trans. Graph.* 25, 3 (2006), 541–548.

C. Li, K. Zhou, and S. Lin. 2014. Intrinsic Face Image Decomposition with Human Face Priors. In *ECCV (5)'14*. 218–233.

H. Li, L. Trutoiu, K. Olszewski, L. Wei, T. Trutna, P.-L. Hsieh, A. Nicholls, and C. Ma. 2015. Facial Performance Sensing Head-Mounted Display. *ACM Trans. Graph.* 34, 4 (July 2015).

Y. Li, S. Liu, J. Yang, and M.-H. Yang. 2017. Generative Face Completion. In *Proc. CVPR*.

C. Liu, H.-Y. Shum, and W. T. Freeman. 2007. Face Hallucination: Theory and Practice. *Int. J. Comput. Vision* 75, 1 (2007), 115–134.

F. Liu, D. Zeng, J. Li, and Q.-j. Zhao. 2017. On 3D face reconstruction via cascaded regression in shape space. *Frontiers of Information Technology & Electronic Engineering* 18, 12 (2017), 1978–1990.

Z. Liu, P. Luo, X. Wang, and X. Tang. 2015. Deep Learning Face Attributes in the Wild. In *IEEE ICCV*.

D. S. Ma, J. Correll, and B. Wittenbrink. 2015. The Chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods* 47, 4 (2015), 1122–1135.

W.-C. Ma, T. Hawkins, P. Peers, C.-F. Chabert, M. Weiss, and P. Debevec. 2007a. Rapid Acquisition of Specular and Diffuse Normal Maps from Polarized Spherical Gradient Illumination. In *Proceedings of the 18th Eurographics Conference on Rendering Techniques (EGSR'07)*. Eurographics Association, Aire-la-Ville, Switzerland, Switzerland, 183–194.

W.-C. Ma, T. Hawkins, P. Peers, C.-F. Chabert, M. Weiss, and P. Debevec. 2007b. Rapid Acquisition of Specular and Diffuse Normal Maps from Polarized Spherical Gradient Illumination. In *Eurographics Symposium on Rendering*.

W.-C. Ma, A. Jones, J.-Y. Chiang, T. Hawkins, S. Frederiksen, P. Peers, M. Vukovic, M. Ouhyoung, and P. Debevec. 2008. Facial Performance Synthesis Using Deformation-driven Polynomial Displacement Maps. In *Proc. SIGGRAPH*. ACM, Article 121, 10 pages.

I. Matthews and S. Baker. 2004. Active Appearance Models Revisited. *Int. J. Comput. Vision* 60, 2 (2004), 135–164.

S. McDonagh, M. Klaudiny, D. Bradley, T. Beeler, I. Matthews, and K. Mitchell. 2016. Synthetic prior design for real-time face tracking. In *3D Vision (3DV), 2016 Fourth International Conference on*. IEEE, 639–648.

U. Mohammed, S. J. D. Prince, and J. Kautz. 2009. Visio-lization: Generating Novel Facial Images. In *ACM Trans. Graph.* ACM, Article 57, 8 pages.

K. Nagano, G. Fyffe, O. Alexander, J. Barbič, H. Li, A. Ghosh, and P. Debevec. 2015. Skin Microstructure Deformation with Displacement Map Convolution. *ACM Trans. Graph.* 34, 4 (2015).

C. Nhan Duong, K. Luu, K. Gia Quach, and T. D. Bui. 2015. Beyond principal components: Deep boltzmann machines for face modeling. In *Proc. CVPR*. 4786–4794.

K. Olszewski, Z. Li, C. Yang, Y. Zhou, R. Yu, Z. Huang, S. Xiang, S. Saito, P. Kohli, and H. Li. 2017. Realistic Dynamic Facial Textures From a Single Image Using GANs. In *IEEE ICCV*.

K. Olszewski, J. J. Lim, S. Saito, and H. Li. 2016. High-Fidelity Facial and Speech Animation for VR HMDs. *ACM Trans. Graph.* 35, 6 (December 2016).

D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. 2016. Context encoders: Feature learning by inpainting. In *Proc. CVPR*. 2536–2544.

A. Radford, L. Metz, and S. Chintala. 2015. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *CoRR* abs/1511.06434 (2015). arXiv:1511.06434

E. Richardson, M. Sela, and R. Kimmel. 2016. 3D face reconstruction by learning from synthetic data. In *3D Vision (3DV), 2016 Fourth International Conference on*. IEEE, 460–469.

E. Richardson, M. Sela, R. Or-El, and R. Kimmel. 2017. Learning detailed face reconstruction from a single image. In *Proc. CVPR*. IEEE, 5553–5562.

S. Romdhani and T. Vetter. 2005. Estimating 3D Shape and Texture Using Pixel Intensity, Edges, Specular Highlights, Texture Constraints and a Prior.. In *Proc. CVPR*. 986–993.

S. Saito, T. Li, and H. Li. 2016. Real-Time Facial Segmentation and Performance Capture from RGB Input. In *ECCV*.

S. Saito, L. Wei, L. Hu, K. Nagano, and H. Li. 2017. Photorealistic Facial Texture Inference Using Deep Neural Networks. In *Proc. CVPR*.

M. Sela, E. Richardson, and R. Kimmel. 2017. Unrestricted facial geometry reconstruction using image-to-image translation. In *IEEE ICCV*. IEEE, 1585–1594.

S. Sengupta, A. Kanazawa, C. D. Castillo, and D. Jacobs. 2017. SfSNet: Learning Shape, Reflectance and Illuminance of Faces in the Wild. *arXiv preprint arXiv:1712.01261* (2017).

F. Shi, H.-T. Wu, X. Tong, and J. Chai. 2014. Automatic acquisition of high-fidelity facial performances using monocular videos. *ACM Trans. Graph.* 33, 6 (2014), 222.

Z. Shu, E. Yumer, S. Hadap, K. Sunkavalli, E. Shechtman, and D. Samaras. 2017. Neural Face Editing with Intrinsic Image Disentangling. *arXiv preprint arXiv:1704.04131* (2017).

Solid Angle. 2016. (2016). http://www.solidangle.com/arnold/.

S. Suwajanakorn, I. Kemelmacher-Shlizerman, and S. M. Seitz. 2014. Total moving face reconstruction. In *ECCV*. Springer, 796–812.

A. Tewari, M. Zollhöfer, P. Garrido, F. Bernard, H. Kim, P. Pérez, and C. Theobalt. 2017a. Self-supervised Multi-level Face Model Learning for Monocular Reconstruction at over 250 Hz. *arXiv preprint arXiv:1712.02859* (2017).

A. Tewari, M. Zollhöfer, H. Kim, P. Garrido, F. Bernard, P. Perez, and C. Theobalt. 2017b. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *IEEE ICCV*, Vol. 2.

The Digital Human League. 2015. Digital Emily 2.0. (2015). http://gl.ict.usc.edu/Research/DigitalEmily2/.

J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. 2016a. Face2Face: Real-time Face Capture and Reenactment of RGB Videos. In *Proc. CVPR*.

J. Thies, M. Zollöfer, M. Stamminger, C. Theobalt, and M. Nießner. 2016b. FaceVR: Real-Time Facial Reenactment and Eye Gaze Control in Virtual Reality. *arXiv preprint arXiv:1610.03151* (2016).

M. Turk and A. Pentland. 1991. Eigenfaces for Recognition. *J. Cognitive Neuroscience* 3, 1 (1991), 71–86.

J. von der Pahlen, J. Jimenez, E. Danvoye, P. Debevec, G. Fyffe, and O. Alexander. 2014. Digital Ira and Beyond: Creating Real-time Photoreal Digital Actors. In *ACM SIGGRAPH 2014 Courses (SIGGRAPH '14)*. ACM, New York, NY, USA, Article 1, 384 pages.

L.-Y. Wei, S. Lefebvre, V. Kwatra, and G. Turk. 2009. State of the art in example-based texture synthesis. In *Eurographics 2009, State of the Art Report, EG-STAR*. Eurographics Association, 93–117.

L.-Y. Wei and M. Levoy. 2000. Fast Texture Synthesis Using Tree-structured Vector Quantization. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '00)*. 479–488.

T. Weyrich, W. Matusik, H. Pfister, B. Bickel, C. Donner, C. Tu, J. McAndless, J. Lee, A. Ngan, H. W. Jensen, and M. Gross. 2006. Analysis of Human Faces using a Measurement-Based Skin Reflectance Model. *ACM Trans. Graph.* 25, 3 (2006), 1013–1024.

C. A. Wilson, A. Ghosh, P. Peers, J.-Y. Chiang, J. Busch, and P. Debevec. 2010. Temporal upsampling of performance geometry using photometric alignment. *ACM Trans. Graph.* 29, 2 (2010), 17.

C. Wu, D. Bradley, M. Gross, and T. Beeler. 2016. An anatomically-constrained local deformation model for monocular face capture. *ACM Trans. Graph.* 35, 4 (2016), 115.

R. A. Yeh*, C. Chen*, T. Y. Lim, S. A. G., M. Hasegawa-Johnson, and M. N. Do. 2017. Semantic Image Inpainting with Deep Generative Models. In *Proc. CVPR*. * equal contribution.

H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. 2017. Pyramid Scene Parsing Network. In *Proc. CVPR*.

J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman. 2017. Toward Multimodal Image-to-Image Translation. In *Advances in Neural Information Processing Systems 30*.

X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li. 2015. High-fidelity pose and expression normalization for face recognition in the wild. In *Proc. CVPR*. 787–796.