Mesoscopic Facial Geometry Inference Using Deep Neural Networks

Loc Huynh[1] Weikai Chen[1] Shunsuke Saito[1,2,3] Jun Xing[1] Koki Nagano[3] Andrew Jones[1]
Paul Debevec[1] Hao Li[1,2,3]

[1]USC Institute for Creative Technologies [2]University of Southern California [3]Pinscreen

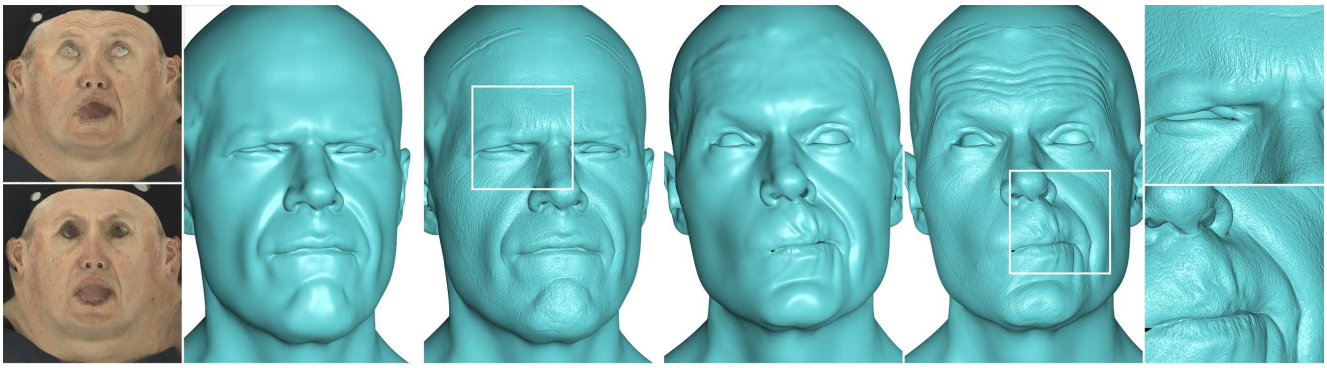


Figure 1: Given a flat-lit facial input textures and a base mesh, our system can synthesize high-resolution facial geometry.

Abstract

We present a learning-based approach for synthesizing facial geometry at medium and fine scales from diffusely-lit facial texture maps. When applied to an image sequence, the synthesized detail is temporally coherent. Unlike current state-of-the-art methods [17, 5], which assume "dark is deep", our model is trained with measured facial detail collected using polarized gradient illumination in a Light Stage [20]. This enables us to produce plausible facial detail across the entire face, including where previous approaches may incorrectly interpret dark features as concavities such as at moles, hair stubble, and occluded pores. Instead of directly inferring 3D geometry, we propose to encode fine details in high-resolution displacement maps which are learned through a hybrid network adopting the state-of-the-art image-to-image translation network [29] and super resolution network [43]. To effectively capture geometric detail at both mid- and high frequencies, we factorize the learning into two separate sub-networks, enabling the full range of facial detail to be modeled. Results from our learning-based approach compare favorably with a high-quality active facial scanhening technique, and require only a single passive lighting condition without a complex scanning setup.

1. Introduction

There is a growing demand for realistic, animated human avatars for interactive digital communication in augmented and virtual reality, but most real-time computer generated humans continue to be simplistic and stylized or require a great deal of effort to construct. An important part of creating a realistic, relatable human avatar is skin detail, from the dynamic wrinkles that form around the eyes, mouth, and forehead that help express emotion, to the fine-scale texture of fine creases and pores that make the skin surface look like that of a real human. Constructing such details on a digital character can take weeks of effort by digital artists, and often employs specialized and expensive 3D scanning equipment to measure skin details from real people. And the problem is made much more complicated by the fact that such skin details are dynamic: wrinkles form and disappear, skin pores stretch and shear, and every change provides a cue to the avatar's expression and their realism.

Scanning the overall shape of a face to an accuracy of a millimeter or two has been possible since the 1980's using commercial laser scanners such as a Cyberware system. In recent years, advances in multiview stereo algorithms such as [14, 15] have enabled facial scanning using passive multiview stereo which can be done with an ordinary setup of consumer digtital cameras. However, recording submillimeter detail at the level of skin pores and fine creases necessary for photorealism remains a challenge. Some of today's best results are obtained in a professional studio capture setup with specialized lighting patterns, such as the polarized gradient photometric stereo process of [34, 20]. Other techniques [25, 22, 24] use high-resolution measurements or statistics of a few skin patches and perform texture synthesis over the rest of the face to imagine what the high-

resolution surface detail might be like. Other work uses a heuristic "dark is deep" shape-from-shading approach [5, 17] to infer geometric surface detail from diffuse texture maps, but can confuse surface albedo variation with geometric structure.

In this work, we endeavor to efficiently reconstruct dynamic medium- and fine-scale geometric facial detail for static facial scans and dynamic facial performances across a wide range of expressions, ages, gender, and skin types without requiring specialized capture hardware. To do this, we propose the first deep learning based approach to infer temporally coherent high-fidelity facial geometry down to the level of skin pore detail directly from a sequence of diffuse texture maps. To learn this mapping, we leverage a database of facial scans recorded with a state-of-the-art active illumination facial scanning system which includes pairs of diffusely-lit facial texture maps and high-resolution skin displacement maps. We then train a convolutional neural network to infer high-resolution displacement maps from the diffuse texture maps, the latter of which can be recorded much more easily with a passive multiview stereo setup. Our hybrid network fuses two components: 1) an image-to-image translation net that translates input texture map to displacement map, and 2) a super-resolution net that generates the high-resolution output given the outcome of preceding network. Our preliminary experiments demonstrate that medium scale and pore-level geometries are encoded in different dynamic ranges. Therefore, we introduce two sub-networks in the image translation net to decouple the learning of middle and high-frequency details. Experimental results indicate our architecture is capable of inferring a full range of detailed geometries with quality that is on par with state-of-the-art facial scanning data.

Compared with conventional methods, our proposed approach provides much faster reconstruction of fine-scale facial geometry thanks to the deep learning framework. In addition, our model is free from certain artifacts which can be introduced using a "dark is deep" prior to infer gemoetric facial detail. Since our model is trained with high-resolution surface measurements from the active illumination scanning system, the network learns the relationship between facial texture maps and geometric detail which is not a simple function of local surface color variation.

The contributions of this work include:

- We present the first deep learning framework that reconstructs high resolution dynamic displacement comparable to active illumination scanning systems entirely from a passive multiview imagery.
- We show how it is possible to learn such inference from sparse but high-resolution geometry data using a two-level image translation network with a conditional GAN combined with a patch-based super resolution network.
- We provide robust reconstruction of both medium and high frequency structures including moles and stubble hair, correctly distinguishing surface pigmentation and the actual surface bumps, outperforming other methods based on high-frequency hallucination or simulation methods.

2. Related Work

Facial Geometry and Appearance Capture. The foundational work of Blanz and Vetter [8] showed that a morphable principal components model built from 3D facial scans can be used to reconstruct a wide variety of facial shapes and overall skin coloration. However, the scans used in their work did not include submillimeter-resolution skin surface geometry, and the nonlinear nature of skin texture deformation would be difficult to embody using such a linear model. For the skin detail synthesis, Saito et al. [40] presented a photorealistic texture inference technique using a deep neural network-based feature correlation analysis. The learned skin texture details can be used to enhance the fidelity of personalized avatars [26]. While the method learns to synthesize mesoscopic facial albedo details, the same approach cannot be trivially extended to the geometry inference since the feature correlation analysis requires a number of geometry scans.

Mesoscopic Facial Detail Capture. There are many techniques for deriving 3D models of faces, and some are able to measure high-resolution skin detail to a tenth of a millimeter, the level of resolution we address in this work. Some of the best results are derived by scanning facial casts from a molding process [27, 1], but the process is time consuming and impossible to apply to a dynamic facial performance. Multi-view stereo provides the basic geometry estimation technique for many facial scanning systems [14, 15, 5, 9, 6, 47, 16, 17]. However, stereo alone typically recovers limited surface detail due to the semi-translucent nature of the skin.

Inferring local surface detail from shape from shading is a well established technique for unconstrained geometry capture [32, 21, 4], and has been employed in digitizing human faces [31, 18, 42, 44, 33, 28, 19]. However, the fidelity of the inferred detail is limited due to the input image captured under unconstrained setting. Beeler et al. [5, 6] applied shape from shading to emboss high-frequency skin shading as hallucinated mesoscopic geometric details for skin pores and creases. While the result is visually plausible, some convexities on the surface can be misclassified as geometric concavities, producing incorrect surface details. More recent work [17] extended this scheme to employ relative shading change ("darker is deeper" heuristics) to mitigate the ambiguity between the dark skin pigmentation and the actual geometric details.

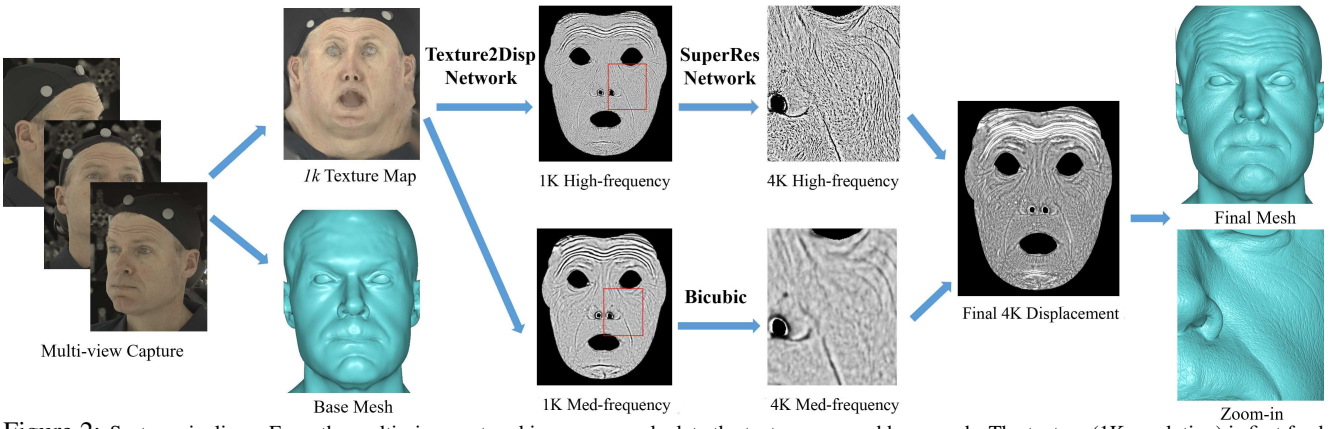


Figure 2: System pipeline. From the multi-view captured images, we calculate the texture map and base mesh. The texture (1K resolution) is first feed into our trained Texture2Disp network to produce a 1K-high and 1K-middle frequency displacement maps, followed by up-sampling them to 4K resolution using our trained SuperRes Network and bicubic interpolation, respectively. The combined 4K displacement map can be embossed to the base mesh to produce the final high detailed mesh.

Active photometric stereo based from specular reflections has been used to measuring detailed surface geometry in devices such as a light stage [12, 34, 24, 20] which can be used to create photoreal digital humans [2, 48, 45]. A variant of the system has been introduced by Weyrich et al. [49] for statistical photorealistic reflectance capture. Ghosh et al. [20] employed multi-view polarized spherical gradient illumination to estimate sub-milimeter accurate mesoscopic displacements. Static [24] and dynamic [36] microstructures can be recorded using a similar photometric system or a contact based method [25, 30]. Photometric stereo techniques have been extended to video performance capture [35, 50, 23]. However, these require high-speed imaging equipment and synchronized active illumination to record the data.

Geometric Detail Inference. Previous work has successfully employed data-driven approaches for inferring facial geometric details. Skin detail can be synthesized using data-driven texture synthesis [25] or statistical skin detail models [22]. Dynamic facial details can be inferred from sparse deformation using polynomial texture maps[35] or radial basis functions [7]. However, these methods can require significant effort to apply to a new person. More recently Cao et al [10] proposed to locally regress medium-scale details (e.g. expression wrinkles) from high-resolution capture data. While generalizing to new test data after training, their approach cannot capture pore-level details.

A neural network-based approach has been introduced for predicting image-domain pixel-wise transformation with a conditional GAN [29] and inference of surface normals for general objects [3, 11]. For facial geometry inference, Trigeorgis et al. [46] employed fully convolutional networks to infer a coarse face geometry through surface normal estimation. More recently, Richardson et al. [38] and Sela et al. [41] presented a learning-based approach to reconstruct detailed facial geometry from a single image. However, none of the previous works has addressed the inference of mesoscopic facial geometry, perhaps due the limited availability of high fidelity geometric data.

3. Overview

Figure 2 illustrates the pipeline of our system. In the pre-processing, we first reconstruct a base face mesh and a 1K-resolution UV texture map from input multi-view images of a variety of subjects and expressions by fitting a template model with consistent topology using the state of the art dynamic facial reconstruction [17]. The texture map is captured under an uniformly lit environment to mimic the natural lighting. Our learning framework takes a texture map as an input and generates a high-quality 4K-resolution displacement map that encodes a full range of geometric details. In particular, it consists of two major components: a two-level image-to-image translation network that synthesizes 1K resolution medium and high frequency displacement maps from the input facial textures, and a patch-based super resolution network that enhances the high frequency displacement map to 4K resolution, introducing sub-pore level details. The medium frequency displacement map is upsampled using a naive bicubic upsampling, which turns out to be sufficient in our experiments. The final displacement is obtained by combining individually inferred medium and high frequency displacement maps. Finally the inferred displacement is applied on the given base mesh to get the final geometry with fine-scale details.

4. Geometry Detail Separation

A key to the success of our method is carefully processed geometric data and its representation. While recent research has directly trained neural networks with 3D vertex positions [37], these approaches can be memory intensive. While unstructured representation is suitable for gen-

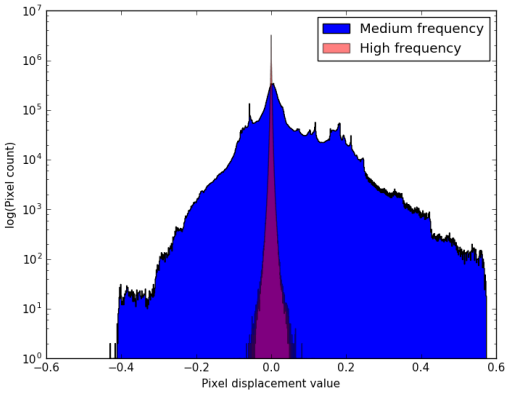
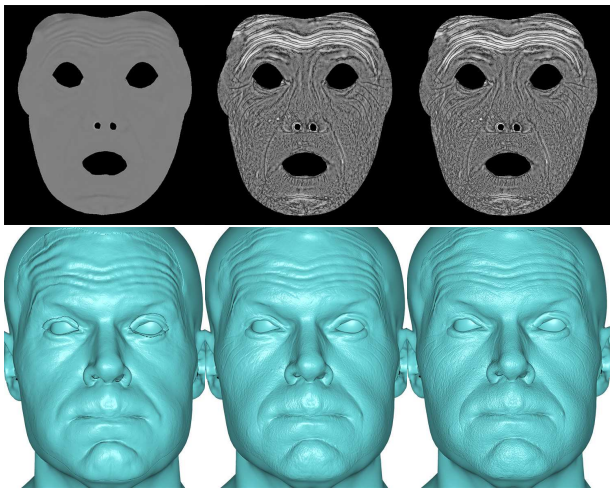


Figure 3: The histogram of the medium and high frequency pixel count shows that the majority of the high frequency details lie within a very narrow band of displacement values compared to the medium frequency values spreading over a broader dynamic range.

eral objects, it can be suboptimal for human faces, which assumes many common parts. In this work, we encode our facial mesoscopic geometry details in a high resolution displacement map parameterized in a common 2D texture space. The main advantages to use such a representation is two-fold. First, a displacement map is a commonly used [35, 24, 36] and lightweight representation than full 3D coordinates, requiring only a single channel to encode surface details. In particular, human faces deform to develop similar skin texture patterns across different individuals. Graham et al. [24] showed that cross-subject transfer of high frequency geometry details is possible among similar ages and genders. With the displacement data parameterized in a common UV space encoding the same facial regions of different individuals, this helps the network to encapsulate the geometric characteristics of each facial region from a limited number of facial scans. While our method assumes fixed topology, existing methods can also be used to convert between different UV coordinate systems. Secondly, from a learning point of view, 2D geometric representation can take advantage of recent advances in convolutional neural networks that could serve for our purpose.

Our displacement data encodes high resolution geometry details that are beyond the resolution of a few tens of thousands vertices in our base mesh. Thus it contains relatively large tens of milimeter forehead wrinkles to submilimeter fine details. Figure 3 shows the histogram of the displacement pixel count shown in a log scale with respect to the value of the displacement. As shown here, there is a spike in the histogram within a very small displacement value, implying that there are distinctive geometry features at different scales. Special care must be taken to properly learn such multi-scale nature of facial geometry details. Our experiment shows that, if we naively train our texture-to-displacement network using the unfiltered displacement,



(a) no separation, 1K (b) separation, 1K (c) separation, 4K

Figure 4: By separating the displacement map into high and middle frequencies, the network could learn both the dominant structure and subtle details (a and b), and the details could be further enhanced via super-resolution (b and c).

the medium-scale geometry dominates the dynamic range of pixel value and leaves the high frequency details trivial, making the network unable to learn high frequency geometric details. Inspired by previous work [35], we factor the displacement into the medium and high frequency components, and learn them individually via two subnetworks. In particular, during training, we decouple the ground truth displacement map $\mathcal{D}$ into two component maps – $\mathcal{D}_L$ and $\mathcal{D}_H$, which capture the medium and high frequency geometry respectively. The resulting data is fed into the corresponding subnetworks of the Image-to-Image translation network.

We show in Figure 4 that the geometry detail separation is the key to achieve faithful reconstruction that capture multi-scale facial details (Figure 4b), while a naive approach without decoupling tends to lose all high frequency details, introducing artifacts (Figure 4a).

5. Texture-to-Displacement Network

Human skin texture provides a great deal of perceptual depth information through skin shading, and previous work has leveraged the apparent surface shading to reveal the underlying surface geometry from a variant of models and heuristics. The inference of the truthful surface details is non-trivial due to the complex light transport in human skin and non-linear nature of skin deformation. To mitigate some of these challenges, we employ uniformly lit texture as an input to our system. Since we employ the input texture and the displacement maps registered at a common pixel coordinate, our inference problem can be posed as image-space translation problem. In this paper, we propose to directly learn this image-to-image translation by leveraging the pairs of input texture maps and corresponding geometry

encoded in the displacement map. To our knowledge, we are the first to solve pore-level geometry inference from an input texture as image-to-image translation problem.

We adopt the state-of-the-art image-to-image translation network using a conditional GAN and U-net architecture with skip connections [29]. The advantage of the proposed network is three-fold. First, the adversarial training facilitates the learning of the input modality manifold and produces sharp results, which is essential for high frequency geometry inference. On the other hand, naive pixel-wise reconstruction loss in L2 or L1 norm often generates a blurry output, as demonstrated in [29]. Furthermore, a patch-based discriminator, which makes real/fake decision in local patches using a fully convolutional network, captures local structures in each receptive field. As the discriminator in each patch shares the weights, the network can effectively learn variations of skin details even if the large amount of data is not available. Last but not least, the U-net architecture with skip connections utilizes local details and global features to make inference [39]. Combining a local feature analysis and global reasoning greatly improves the faithful reconstruction especially when underlying skin albedo ambiguates the translation (e.g. skin pigmentation, moles). Our texture-to-displacement network consists of two branches, each fulfilled by the image-to-image translation network. The two subnetworks infer the medium and high frequency displacement maps from the same input texture, respectively.

6. Super-Resolution Network

The effective texture resolution is determined by the ratio of the target face size and the final target resolution (in our work submilimeter details). In our setting, we find that no smaller than a 4K resolution displacement is detailed enough to resolve the pore-level geometry details we want to produce photorealistic rendering. However, in practice applying an image-to-image translation network to a texture more than 1K resolution pixel is computationally demanding and can be beyond the capacity of the modern GPU hardware. To overcome the limitation in the resolution, we propose to further upsample the resulting displacement map using a patch-based super-resolution network. We build our super-resolution network based upon the state-of-the-art super resolution network using sub-pixel convolution [43]. During the training, we downsample the 4K ground-truth displacement maps $\{\mathcal{D}_{hr}\}$ to obtain its corresponding 1K-resolution training set $\{\mathcal{D}_{lr}\}$. We then randomly pick pairs of a 64×64 patch from $\mathcal{D}_{lr}$ and their corresponding 256×256 patch from $\mathcal{D}_{hr}$, which are fed into the network for training. At test time, we first divide the input image into a regular grid, with each of the block forming a 64×64 patch image. We then upsample each patch to 256×256 resolution using the super-resolution network. Finally, to ensure consistency between patch boundaries, we stitch the resulting upsampled patches using image quilting [13] to produce a 4K displacement.

7. Implementation Details

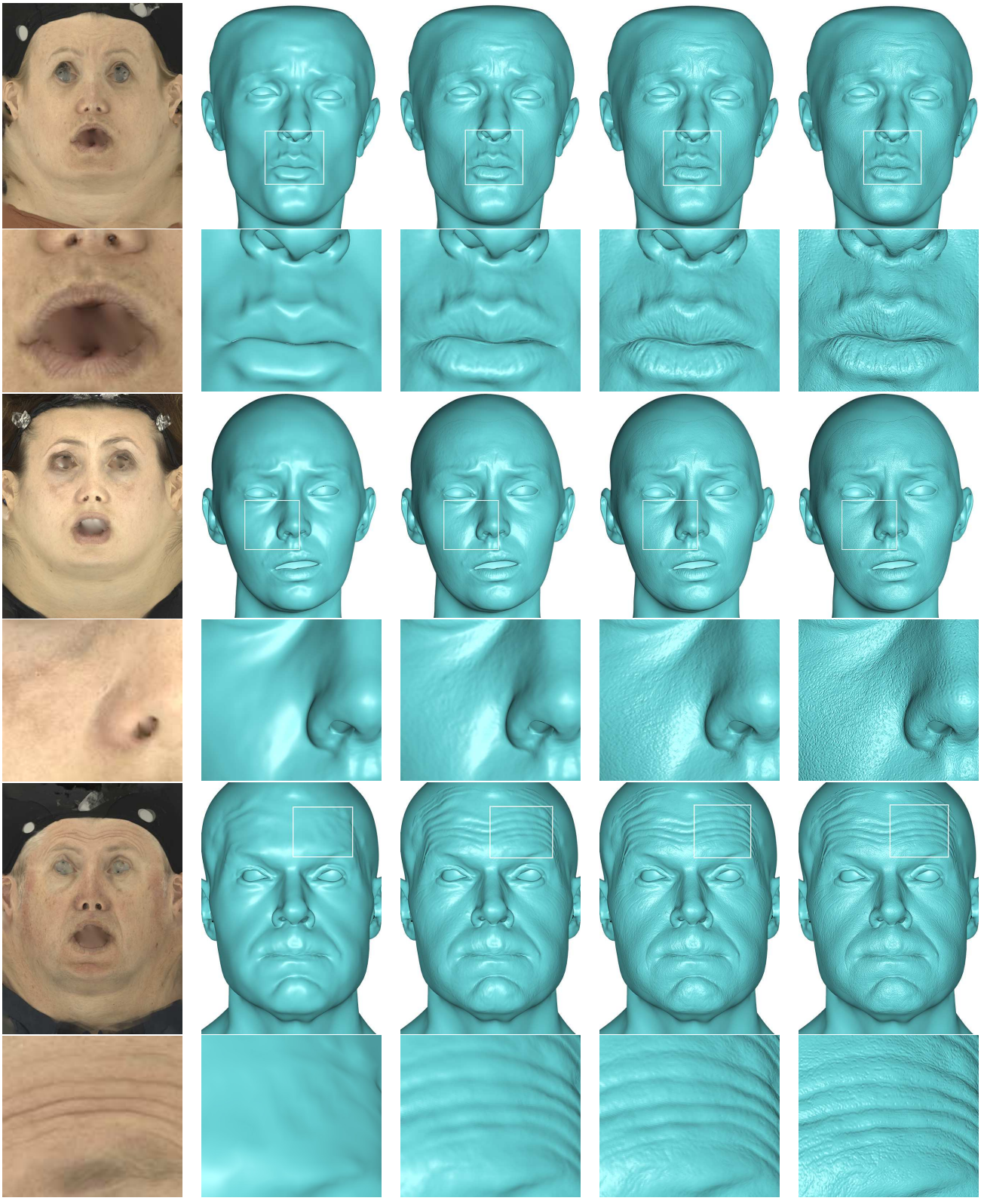
It is important that the training data covers a wide range of ethnic, age, genders, and skin tones. We collected 328 corresponded Light Stage facial scans [20] as ground truth photometric stereo to train the network. This includes 19 unique subjects, between the ages of 11 and 56, with multiple expressions per subject to capture wrinkle and pore dynamics. 6 additional subjects were used to test system performance. For each collected displacement, we apply a Gaussian filter to remove all high frequency data, obtaining the medium frequency displacement map $\mathcal{D}_L$. The high-frequency displacement map can be calculated by subtraction $\mathcal{D}_H = \mathcal{D} - \mathcal{D}_L$. Given the histogram of the displacement maps, we iteratively optimize the filter size of the Gaussian filter so that $\mathcal{D}_H$ covers only high frequency data. We find that the filter size of 29 at 4K resolution gives the best results most of examples. We apply a $\times 64$ scale to the high frequency values to distribute the values well over the limited pixel intensity to facilitate the convergence during learning. For the medium frequency data, which usually exhibits higher displacement range, we apply a sigmoid function so that all the values fit well into the pixel range without clipping. This step takes less than a second for a 1K displacement map.

We train our network with pairs of texture and displacement maps at 1K resolution. The training time on a single NVidia Titan X GPU with 12GB memory is around 8 hours. For the super resolution network, we feed in displacement maps at 4k resolution to train. It takes less than 2 hours to train with the same GPU. At test time, it takes one second to get both 1K resolution displacement maps from a 1K input texture map. Then these maps are up-sampled using our super resolution network. We get the final 4K displacement map after 5 seconds.

8. Experimental Results

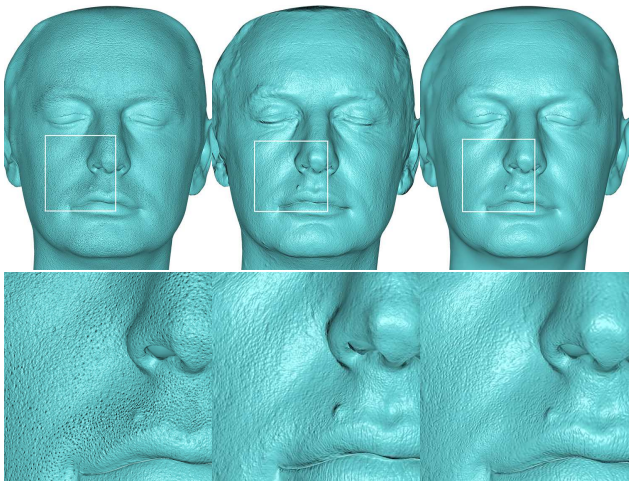
We evaluate the effectiveness of our approach on different input textures with a variety of subjects and expressions. In Figure 5, we show the synthesized geometries embossed by (c) only medium-scale details, (d) 1K and (e) 4K combined multi-scale (both medium and high frequency) displacement maps, with the the input textures and base mesh shown in the first and second column, respectively. As seen from the results, our method can faithfully capture both the medium and fine scale geometries. The final geometry synthesized using the 4K displacement map exhibits meso-scale geometry on par with active facial scanning. None of these subjects are used in training the network, and show the the robustness of our method to a variety of texture qualities, expressions, gender, and ages.

We validate the effectiveness of geometry detail separation by comparing with an alternative solution which does



(a) Input texture (b) Base mesh (c) Med-frequency (d) 1K multi-scale (e) 4K multi-scale

Figure 5: Synthesis results given different input textures with variations in subject identity and expression. From (a) to (e), we show the input texture, base mesh, the output geometry with the medium, 1K multi-scale (both medium and high frequency) and 4K multi-scale frequency displacement map. The closeup is shown under each result.



(a) Beeler et al.[6] (b) Ground truth [20] (c) our method

Figure 6: Inferred detail comparison.

not decouple middle and high frequencies. As illustrated in Figure 4a, the displacement map learned from the alternative method fails to capture almost all the high frequency details while introducing artifacts in middle frequencies, which is manifested in the embossed geometry. Our method, on the other hand, faithfully replicates both medium and fine scale details in the resulting displacement map (Figure 4b).

We also assess the effectiveness of the proposed super-resolution network in our framework. Figure 4c and Figure 4b show the results with and without the super-solution network, respectively. The reconstructed result using super-solution network outperforms its opponent significantly in faithfully replicating mesocopic facial structures.

Comparisons. We compare the reconstruction quality of our method with Beeler et al. [5] and the ground truth by Ghosh et al. [20]. As demonstrated in Figure 6, our reconstruction (right) generally agrees with the ground truth (middle) in capturing the roughness variation between the tip of the nose and the cheek region, and the mole by the upper lip. The "dark is deep" heuristic [5] (left), on the other hand, fails to capture these geometric differences. In Figure 7, we provide the quantitative evaluation comparing with Beeler et al. [5]. We measure the reconstruction error using the L_1 metric between ours and the ground truth displacement map provided by Ghosh et al. [20]. The resulting error map is visualized in false color with red and blue indicating the absolute max difference 1 mm to 0 mm, respectively. As manifested in Figure 7, Beeler et al. [5] is prone to introduce larger reconstruction error particularly for regions with stubble hair and eyebrows. Our model, trained with photometric scans, achieves superior accuracy and robust inference without being confused too much by the local skin albedo variations.

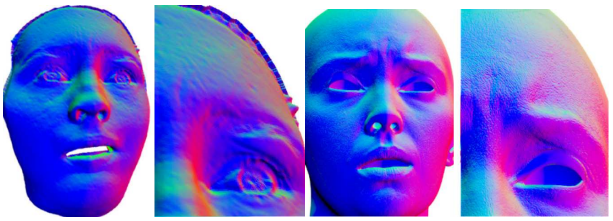
Our system can also generate dynamic displacement

(a) Light Stage (b) our method (c) dark is deep

Figure 7: High frequency details of our method (center) comparison with ground truth Light Stage data [20] (left) and "dark is deep" heuristic.[5] (right)

maps for video performances. In the supplemental video, we demonstrate that our results are stable across frames and accurately represent changing wrinkles and pores. We also compare our results against a dynamic sequence with the state of the art dynamic multi-view face capture of Fyffe et al. 2017 [17]. Fyffe et al relies on multiview stereo to reconstruct medium frequencies and on inter-frame changes in shading to infer high-freqency detail. Our method produces more accurate fine-scale details as it is trained on photometric stereo and can be computed on each frame independently.

We also evaluate our technique against similar neural network synthesis methods. Sela et al [41] use an image-to-image translation network but to infer a facial depth map and dense correspondences from a single image. This is followed by non-rigid registration and shape from shading similar to Beeler et al. [5]. Their generated image lacks fine-scale details as these are not encoded in their network (see Figure 8). Bansal et al. [3] offers the state-of-the-art performance on estimating surface normal using convolu-



(a) Sela et al. [41] (b) our method

Figure 8: Compared with Sela et al. [41], our method could produce more detailed normal map.

tional neural network. We also provide the comparison with Bansal et al. [3] in terms of normal prediction accuracy in the supplementary material.

User Study. We assessed the realism of our inferred geometry by a user study. Users are asked to sort the renderings of 3D face without skin textures from unrealistic to realistic. We used 6 subjects for rendering using (1) our synthesized geometry, (2) the Light Stage [20], and (3) the "dark is deep" synthesis [5] and randomly sorted them, aligning in the same head orientation to remove bias. We collected 58 answers from 25 subjects. 20.7% of users think our reconstructions are the most realistic, while 67.2% and 12.1% of people find the Light Stage and [5] more realistic. Although the Light Stage still shows superior performance in terms of realism, our method is favorably compared with the geometry synthesis method [5].

9. Discussion and Future Work

Our primary observation is that a high-resolution diffuse texture map contains enough implicit information to accurately infer useful geometric detail. Secondly, neural network based synthesis trained on ground truth photometric stereo data outperforms previous shape from shading heuristics. Our system can successfully differentiate between skin pores, stubble, wrinkles, and moles based on their location on the face, and how their appearance changes across different subjects and expressions. Our method generates stable high-resolution displacement maps in only a few seconds, with realistic dynamics suitable for both static scans and video performances.

The limitation of our method is that the training data need to be carefully corresponded. However, our learning framework does not strictly require dense registration since there is no meaningful pore-to-pore correspondence across different identities. We ensure in the training that correspondence is roughly maintained in UV space across different subjects so that the generated displacement maintains the correct skin detail distributions. Though our training data was captured using flat-lit environment, our method could be integrated with previous albedo synthesis techniques which compensate for varying illumination and fill in occluded regions [40] in order to infer facial details of unconstrained images in the wild. We show additional results

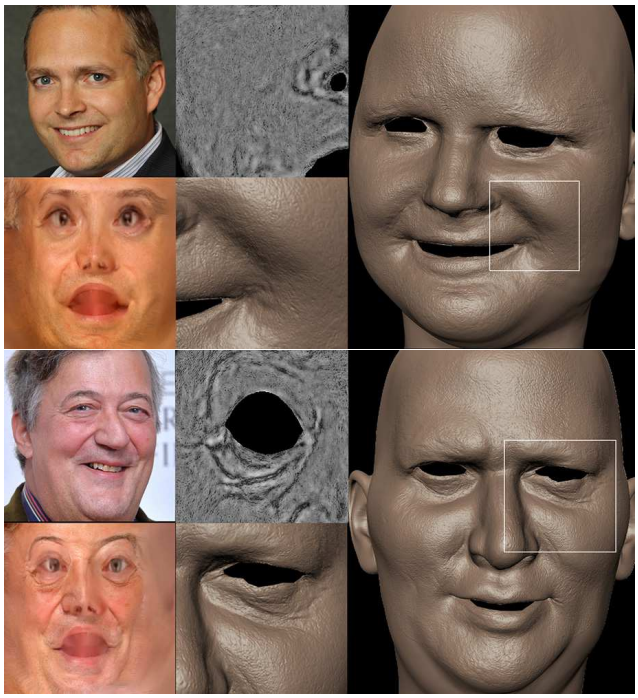


Figure 9: Results with unconstrained images. Left: input image, texture. Middle: displacement (zoom-in), rendering (zoom-in). Right: rendering

in Figure 9 to support these claims using a novel topology obtained from a conventional 3D morphable model. While our training dataset contains several examples of commonly applied cosmetics, more pronounced theatrical makeup may introduce displacement artifacts (see the supplementary material for a failure example). We believe our results will continue to improve with additional training data featuring unusual moles, blemishes, and scars. We would also like to incorporate other channels of input. For example, wrinkles are correlated with low-frequency geometry stress [35, 7] and local specular highlights can provide additional detail information.

Acknowledgements We wish to thank all of our actors, especially Mike Seymour, Emily O'Brien, Ronald Mallet, Tony Tung, and Gallaudet University for giving us permissions to use their scans. This research is supported by Adobe, Sony, the Google Faculty Research Award, the Okawa Foundation Research Grant, the Andrew and Erna Viterbi Early Career Chair, the Office of Naval Research (ONR), under award number N00014-15-1-2639, and the U.S. Army Research Laboratory (ARL) under contract W911NF-14-D-0005. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of ONR, ARL, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purpose notwithstanding any copyright annotation thereon.

References

[1] G. Acevedo, S. Nevshupov, J. Cowely, and K. Norris. An accurate method for acquiring high resolution skin displacement maps. In *ACM SIGGRAPH 2010 Talks*, SIGGRAPH '10, pages 4:1–4:1, New York, NY, USA, 2010. ACM.

[2] O. Alexander, M. Rogers, W. Lambeth, M. Chiang, and P. Debevec. The digital emily project: Photoreal facial modeling and animation. In *ACM SIGGRAPH 2009 Courses*, SIGGRAPH '09, pages 12:1–12:15, New York, NY, USA, 2009. ACM.

[3] A. Bansal, B. Russell, and A. Gupta. Marr revisited: 2d-3d alignment via surface normal prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5965–5974, 2016.

[4] J. T. Barron and J. Malik. Shape, illumination, and reflectance from shading. *TPAMI*, 2015.

[5] T. Beeler, B. Bickel, P. Beardsley, B. Sumner, and M. Gross. High-quality single-shot capture of facial geometry. *ACM Trans. on Graphics (Proc. SIGGRAPH)*, 29(3):40:1–40:9, 2010.

[6] T. Beeler, F. Hahn, D. Bradley, B. Bickel, P. Beardsley, C. Gotsman, R. W. Sumner, and M. Gross. High-quality passive facial performance capture using anchor frames. In *ACM Transactions on Graphics (TOG)*, volume 30, page 75. ACM, 2011.

[7] B. Bickel, M. Lang, M. Botsch, M. A. Otaduy, and M. Gross. Pose-space animation and transfer of facial details. In *Proceedings of the 2008 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, SCA '08, pages 57–66, Aire-la-Ville, Switzerland, Switzerland, 2008. Eurographics Association.

[8] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194. ACM Press/Addison-Wesley Publishing Co., 1999.

[9] D. Bradley, W. Heidrich, T. Popa, and A. Sheffer. High resolution passive facial performance capture. In *ACM transactions on graphics (TOG)*, volume 29, page 41. ACM, 2010.

[10] C. Cao, D. Bradley, K. Zhou, and T. Beeler. Real-time high-fidelity facial performance capture. *ACM Transactions on Graphics (TOG)*, 34(4):46, 2015.

[11] W. Chen, D. Xiang, and J. Deng. Surface normals in the wild. *arXiv preprint arXiv:1704.02956*, 2017.

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

[13] A. A. Efros and W. T. Freeman. Image quilting for texture synthesis and transfer. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 341–346. ACM, 2001.

[14] Y. Furukawa and J. Ponce. Dense 3d motion capture for human faces. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1674–1681. IEEE, 2009.

[15] Y. Furukawa and J. Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence*, 32(8):1362–1376, 2010.

[16] G. Fyffe, A. Jones, O. Alexander, R. Ichikari, and P. Debevec. Driving high-resolution facial scans with video performance capture. *ACM Trans. Graph.*, 34(1):8:1–8:14, Dec. 2014.

[17] G. Fyffe, K. Nagano, L. Huynh, S. Saito, J. Busch, A. Jones, H. Li, and P. Debevec. Multi-view stereo on consistent face topology. In *Computer Graphics Forum*, volume 36, pages 295–309. Wiley Online Library, 2017.

[18] P. Garrido, L. Valgaerts, C. Wu, and C. Theobalt. Reconstructing detailed dynamic face geometry from monocular video. In *ACM Trans. Graph. (Proceedings of SIGGRAPH Asia 2013)*, volume 32, pages 158:1–158:10, November 2013.

[19] P. Garrido, M. Zollhöfer, D. Casas, L. Valgaerts, K. Varanasi, P. Perez, and C. Theobalt. Reconstruction of personalized 3d face rigs from monocular video. *ACM Trans. Graph. (Presented at SIGGRAPH 2016)*, 35(3):28:1–28:15, 2016.

[20] A. Ghosh, G. Fyffe, B. Tunwattanapong, J. Busch, X. Yu, and P. Debevec. Multiview face capture using polarized spherical gradient illumination. *ACM Trans. Graph.*, 30(6):129:1–129:10, 2011.

[21] M. Glencross, G. J. Ward, F. Melendez, C. Jay, J. Liu, and R. Hubbold. A perceptually validated model for surface depth hallucination. In *ACM SIGGRAPH 2008 Papers*, SIGGRAPH '08, pages 59:1–59:8, New York, NY, USA, 2008. ACM.

[22] A. Golovinskiy, W. Matusik, H. Pfister, S. Rusinkiewicz, and T. Funkhouser. A statistical model for synthesis of detailed facial geometry. *ACM Trans. Graph.*, 25(3):1025–1034, 2006.

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

[24] P. Graham, B. Tunwattanapong, J. Busch, X. Yu, A. Jones, P. Debevec, and A. Ghosh. Measurement-based synthesis of facial microgeometry. In *Computer Graphics Forum*, volume 32, pages 335–344. Wiley Online Library, 2013.

[25] A. Haro, B. Guenter, and I. Essa. Real-time, Photo-realistic, Physically Based Rendering of Fine Scale Human Skin Structure. In S. J. Gortle and K. Myszkowski, editors, *Eurographics Workshop on Rendering*, 2001.

[26] L. Hu, S. Saito, L. Wei, K. Nagano, J. Seo, J. Fursund, I. Sadeghi, C. Sun, Y.-C. Chen, and H. Li. Avatar digitization from a single image for real-time rendering. In *ACM Trans. Graph. (Proceedings of SIGGRAPH Asia 2017)*, number 10, November 2017.

[27] W. Hyneman, H. Itokazu, L. Williams, and X. Zhao. Human face project. In *ACM SIGGRAPH 2005 Courses*, SIGGRAPH '05, New York, NY, USA, 2005. ACM.

[28] A. E. Ichim, S. Bouaziz, and M. Pauly. Dynamic 3d avatar creation from hand-held video input. *ACM Trans. Graph.*, 34(4):45:1–45:14, 2015.

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

[30] M. K. Johnson, F. Cole, A. Raj, and E. H. Adelson. Microgeometry capture using an elastomeric sensor. *ACM Transactions on Graphics (Proc. ACM SIGGRAPH)*, 30(4):46:1–46:8, 2011.

[31] I. Kemelmacher-Shlizerman and R. Basri. 3d face reconstruction from a single image using a single reference face shape. *IEEE TPAMI*, 33(2):394–405, 2011.

[32] M. S. Langer and S. W. Zucker. Shape-from-shading on a cloudy day. *J. Opt. Soc. Am. A*, 11(2):467–478, Feb 1994.

[33] C. Li, K. Zhou, and S. Lin. Intrinsic face image decomposition with human face priors. In *ECCV (5)'14*, pages 218–233, 2014.

[34] W.-C. Ma, T. Hawkins, P. Peers, C.-F. Chabert, M. Weiss, and P. Debevec. 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, pages 183–194, Aire-la-Ville, Switzerland, Switzerland, 2007. Eurographics Association.

[35] W.-C. Ma, A. Jones, J.-Y. Chiang, T. Hawkins, S. Frederiksen, P. Peers, M. Vukovic, M. Ouhyoung, and P. Debevec. Facial performance synthesis using deformation-driven polynomial displacement maps. In *ACM SIGGRAPH Asia 2008 Papers*, SIGGRAPH Asia '08, pages 121:1–121:10, New York, NY, USA, 2008. ACM.

[36] K. Nagano, G. Fyffe, O. Alexander, J. Barbič, H. Li, A. Ghosh, and P. Debevec. Skin microstructure deformation with displacement map convolution. *ACM Transactions on Graphics (Proceedings SIGGRAPH 2015)*, 34(4), 2015.

[37] C. R. Qi, H. Su, K. Mo, and L. J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *CoRR*, abs/1612.00593, 2016.

[38] E. Richardson, M. Sela, R. Or-El, and R. Kimmel. Learning detailed face reconstruction from a single image. *arXiv preprint arXiv:1611.05053*, 2016.

[39] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

[40] S. Saito, L. Wei, L. Hu, K. Nagano, and H. Li. Photorealistic facial texture inference using deep neural networks. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017.

[41] M. Sela, E. Richardson, and R. Kimmel. Unrestricted facial geometry reconstruction using image-to-image translation. *arXiv preprint arXiv:1703.10131*, 2017.

[42] F. Shi, H.-T. Wu, X. Tong, and J. Chai. Automatic acquisition of high-fidelity facial performances using monocular videos. *ACM Trans. Graph.*, 33(6):222:1–222:13, 2014.

[43] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1874–1883, 2016.

[44] S. Suwajanakorn, I. Kemelmacher-Shlizerman, and S. M. Seitz. *Total Moving Face Reconstruction*, pages 796–812. Springer International Publishing, Cham, 2014.

[45] The Digital Human League. Digital Emily 2.0, 2015. http://gl.ict.usc.edu/Research/DigitalEmily2/.

[46] G. Trigeorgis, P. Snape, I. Kokkinos, and S. Zafeiriou. Face normals ?in-the-wild? using fully convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[47] L. Valgaerts, C. Wu, A. Bruhn, H.-P. Seidel, and C. Theobalt. Lightweight binocular facial performance capture under uncontrolled lighting. In *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia 2012)*, volume 31, pages 187:1–187:11, November 2012.

[48] J. von der Pahlen, J. Jimenez, E. Danvoye, P. Debevec, G. Fyffe, and O. Alexander. Digital ira and beyond: Creating real-time photoreal digital actors. In *ACM SIGGRAPH 2014 Courses*, SIGGRAPH '14, pages 1:1–1:384, New York, NY, USA, 2014. ACM.

[49] T. Weyrich, W. Matusik, H. Pfister, B. Bickel, C. Donner, C. Tu, J. McAndless, J. Lee, A. Ngan, H. W. Jensen, and M. Gross. Analysis of human faces using a measurement-based skin reflectance model. *ACM Trans. on Graphics (Proc. SIGGRAPH 2006)*, 25(3):1013–1024, 2006.

[50] C. A. Wilson, A. Ghosh, P. Peers, J.-Y. Chiang, J. Busch, and P. Debevec. Temporal upsampling of performance geometry using photometric alignment. *ACM Transactions on Graphics (TOG)*, 29(2):17, 2010.