Supplementary Material:
Unrestricted Facial Geometry Reconstruction Using Image-to-Image Translation

A. Additional Network Details

Here, we summarize additional considerations concerning the network and its training procedure.

- The proposed architecture is based on the one introduced in [6]. For allowing further refinement of the results, three additional convolution layers with a kernel of size 1×1 were concatenated at the end. Following the notations of [6], the encoder architecture is given as

$$C64 - C128 - C256 - C512 - C512 - C512 - C512 - C512,$$

  while the decoder is given by

$$CD512 - CD512 - CD512 - C512 - C512 - C256 - C128 - C64 - C^*64 - C^*32 - C^*4,$$

  where C^* represents a 1×1 convolution with stride 1.

- The resolution of the input and output training images was 512×512 pixels. While this is a relatively large input size for training, the Image-to-Image architecture was able to process it successfully, and provided accurate results. Although, one could train a network on smaller resolutions and then evaluate it on larger images, as shown in [6], we found that our network did not successfully scale up for unseen resolutions.

- While a single network was successfully trained to retrieve both depth and correspondence representations, our experiments show that training separated networks to recover the representations is preferable. Note that the architectures of both networks were identical. This can be justified by the observation that during training, a network allocates its resources for a specific translation task and the representation maps we used have different characteristics.

- A necessary parameter for the registration step is the scale of the face with respect to the image dimensions. While this can be estimated based on global features, such as the distance between the eyes, we opted to retrieve it directly by training the network to predict the x and y coordinates of each pixel in the image alongside the z coordinate.

B. Additional Registration and Refinement Details

Next, we provide a detailed version of the iterative deformation-based registration phase, including implementation details of the fine detail reconstruction.

B.1. Non-Rigid Registration

First, we turn the x,y and z maps from the network into a mesh, by connecting four neighboring pixels, for which the coordinates are known, with a couple of triangles. This step yields a target mesh that might have holes but has dense map to our template model. Based on the correspondence given by the network, we compute the affine transformation from a template face to the mesh. This operation is done by minimizing the squared Euclidean distances between corresponding vertex pairs. To handle outliers, a RANSAC approach is used [4] with $1,000$ iterations and a threshold of 3 millimeters for detecting inliers. Next, similar to [8], an iterative non-rigid registration process deforms the transformed template, aligning it with the mesh. Note, that throughout the registration, only the template is warped, while the target mesh remains fixed. Each iteration involves the following four steps.

1. Each vertex in the template mesh, $v_i \in \mathcal{V}$, is associated with a vertex, c_i, on the target mesh, by evaluating the nearest neighbor in the embedding space. This step is different from the method described in [8], which computes the nearest neighbor in the Euclidean space. As a result, the proposed step allows registering a single template face to different facial identities with arbitrary expressions.

2. Pairs, (v_i, c_i), which are physically distant by more than 1 millimeter and those with normal direction disagreement of more than 5 degrees are detected and ignored in the next step.

3. The template mesh is deformed by minimizing the following energy

$$\begin{aligned} E(V,C) &= \alpha_{p2point} \sum_{(v_i,c_i)\in\mathcal{J}} \|v_i - c_i\|_2^2 \\ &\quad + \alpha_{p2plane} \sum_{(v_i,c_i)\in\mathcal{J}} \left|\vec{n}(c_i)(v_i - c_i)\right|^2 \\ &\quad + \alpha_{memb} \sum_{i\in\mathcal{V}} \sum_{v_j\in\mathcal{N}(v_i)} w_{i,j}\|v_i - v_j\|_2^2, \end{aligned} \tag{1}$$

   where, $w_{i,j}$ is the weight corresponding to the biharmonic Laplacian operator (see [5, 2]), $\vec{n}(c_i)$ is the normal of the corresponding vertex at the target mesh c_i, $\mathcal{J}$ is the set of the remaining associated vertex pairs (v_i, c_i), and $\mathcal{N}(v_i)$ is the set 1-ring neighboring vertices about the vertex v_i. Notice that the first term above is the sum of squared Euclidean distances between matches and its weight $\alpha_{p2point}$ is set to 0.1. The second term is the distance from the point v_i to the tangent plane at the corresponding point on the target mesh, and its weight $\alpha_{p2plane}$ is set to 1. The third term quantifies the stiffness of the mesh and its weight α_{memb} is initialized to 10^8. In practice, the energy term given in Equation 1 is minimized iteratively by an inner loop which contains a linear system of equations. We run this loop until the norm of the difference between the vertex positions of the current iteration and the previous one is below 0.01.

4. If the motion of the template mesh between the current outer iteration and the previous one is below 0.1, we divide the weight α_{memb} by two. This relaxes the stiffness term and allows a greater deformation in the next outer iteration. In addition, we evaluate the difference between the number of remaining pairwise matches in the current iteration versus the previous one. If the difference is below 500, we modify the vertex association step to estimate the physical nearest neighbor vertex, instead of the the nearest neighbor in the space of the embedding given by the network.

This iterative process terminates when the stiffness weight α_{memb} is below 10^6. The resulting output of this phase is a deformed template with fixed triangulation, which contains the overall facial structure recovered by the network, yet, is smoother and complete.

B.2. Fine Detail Reconstruction

Although the network already recovers fine geometric details, such as wrinkles and moles, across parts of the face, a geometric approach can reconstruct details at a finer level, on the entire face, independently of the resolution. Here, we propose an approach motivated by the passive-stereo facial reconstruction method suggested in [1]. The underlying assumption here is that subtle geometric structures can be explained by local variations in the image domain. For some skin tissues, such as nevi, this assumption is inaccurate as the intensity variation results from the albedo. In such cases, the geometric structure would be wrongly modified. Still, for most parts of the face, the reconstructed details are consistent with the actual variations in depth.

The method begins from an interpolated version of the deformed template, provided by a surface subdivision technique. Each vertex $v \in \mathcal{V}_\mathcal{D}$ is painted with the intensity value of the nearest pixel in the image plane. Since we are interested in recovering small details, only the high spatial frequencies, $\mu(v)$, of the texture, $\tau(v)$, are taken into consideration in this phase. For computing this frequency band, we subtract the synthesized low frequencies from the original intensity values. This low-pass filtered part can be computed by convolving the texture with a spatially varying Gaussian kernel in the image domain, as originally proposed. In contrast, since this convolution is equivalent to computing the heat distribution upon the shape after time dt, where the initial heat profile is the original texture, we propose to compute $\mu(v)$ as

$$\mu(v) = \tau(v) - (I - dt \cdot \Delta_g)^{-1}\tau(v), \tag{2}$$

where I is the identity matrix, Δ_g is the cotangent weight discrete Laplacian operator for triangulated meshes [9], and $dt = 0.2$ is a scalar proportional to the cut-off frequency of the filter.

Next, we displace each vertex along its normal direction such that $v' = v + \delta(v)\vec{n}(v)$. The step size of the displacement, $\delta(v)$, is a combination of a data-driven term, $\delta_\mu(v)$, and a regularization one, $\delta_s(v)$. The data-driven term is guided by the high-pass filtered part of the texture, $\mu(v)$. In practice, we require the local differences in the geometry to be proportional to the local variation in the high frequency band of the texture. That is for each vertex v, with a normal $\vec{n}(v)$, and a neighboring vertex v_i, the data-driven term is given by

$$(\mu(v) - \mu(v_i)) = \langle v + \delta_\mu(v)\vec{n}(v) - v_i, \vec{n}(v)\rangle. \tag{3}$$

Thus, the step size assuming a single neighboring vertex can be calculated by

$$\delta_\mu(v) = \gamma(\mu(v) - \mu(v_i)) - \langle v - v_i, \vec{n}(v)\rangle. \tag{4}$$

In the presence of any number of neighboring vertices of v, we compute the weighted average of its 1-ring neighborhood

$$\delta_\mu(v) = \frac{\sum_{v_i \in \mathcal{N}(v)} \alpha(v, v_i)\gamma \left[(\mu(v) - \mu(v_i)) - \langle v - v_i, \vec{n}(v)\rangle\right]}{\sum_{v_i \in \mathcal{N}(v)} \alpha(v, v_i)}, \tag{5}$$

An alternative term can spatially attenuate the contribution of the data-driven term in curved regions for regularizing the reconstruction by

$$\delta_\mu(v) = \frac{\sum\limits_{v_i \in \mathcal{N}(v)} \alpha_{(v,v_i)} \left(\mu(v) - \mu(v_i)\right) \left(1 - \frac{|\langle v - v_i, \vec{n}(v)\rangle|}{\|v - v_i\|}\right)}{\sum\limits_{v_i \in \mathcal{N}(v)} \alpha_{(v,v_i)}}, \tag{6}$$

where $\alpha_{(v,v_i)} = \exp\left(-\|v - v_i\|\right)$. where $\mathcal{N}(v)$ is the set 1-ring neighboring vertices about the vertex v, and $\vec{n}(v)$ is the unit normal at the vertex v.

Since we move each vertex along the normal direction, triangles could intersect each other, particularly in regions with high curvature. To reduce the probability of such collisions, a regularizing displacement field, $\delta_s(v)$, is added. This term is proportional to the mean curvature of the original surface, and is equivalent to a single explicit mesh fairing step [3]. The final surface modification is given by

$$v' = v + (\eta\delta_\mu(v) + (1 - \eta)\delta_s(v)) \cdot \vec{n}(v), \tag{7}$$

for a constant $\eta = 0.2$.

C. Additional Experimental Results

We present additional qualitative results of our method. Figure 1 shows the output representations of the proposed network for a variety of different faces, notice the failure cases presented in the last two rows. One can see that the network generalizes well, but is still limited by the synthetic data. Specifically, the network might fail in presence of occlusions, facial hair or extreme poses. This is also visualized in Figure 2 where the correspondence error is visualized using the texture mapping. Additional reconstruction results of our method are presented in Figure 3. For analyzing the distribution of the error along the face, we present an additional comparison in Figure 4, where the absolute error, given in percents of the ground truth depth, is shown for several facial images.

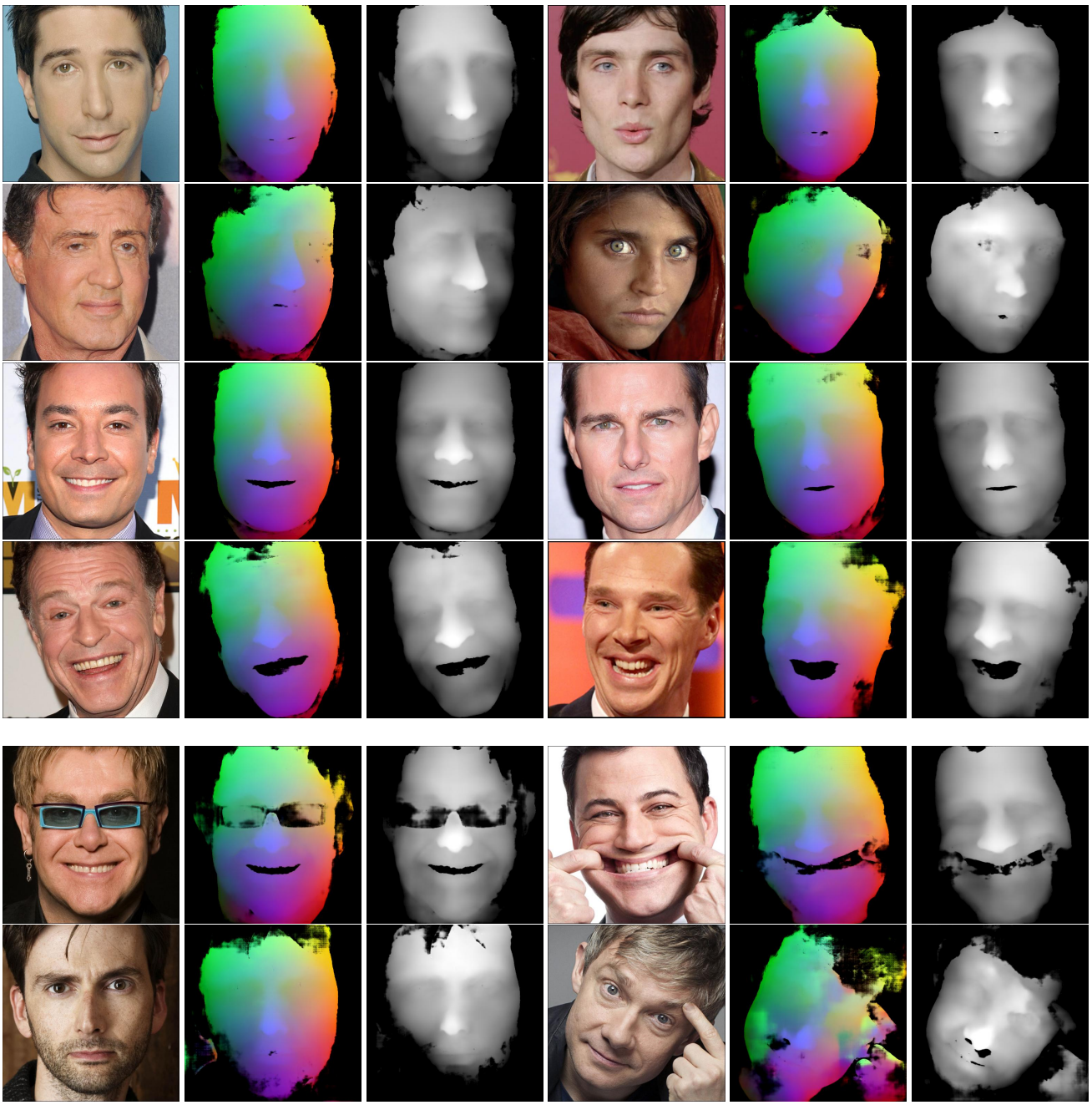

Figure 1: Network Output.

Figure 2: Results under occlusions and rotations. Input images are shown next to the matching correspondence result, visualized using the texture mapping to better show the errors.

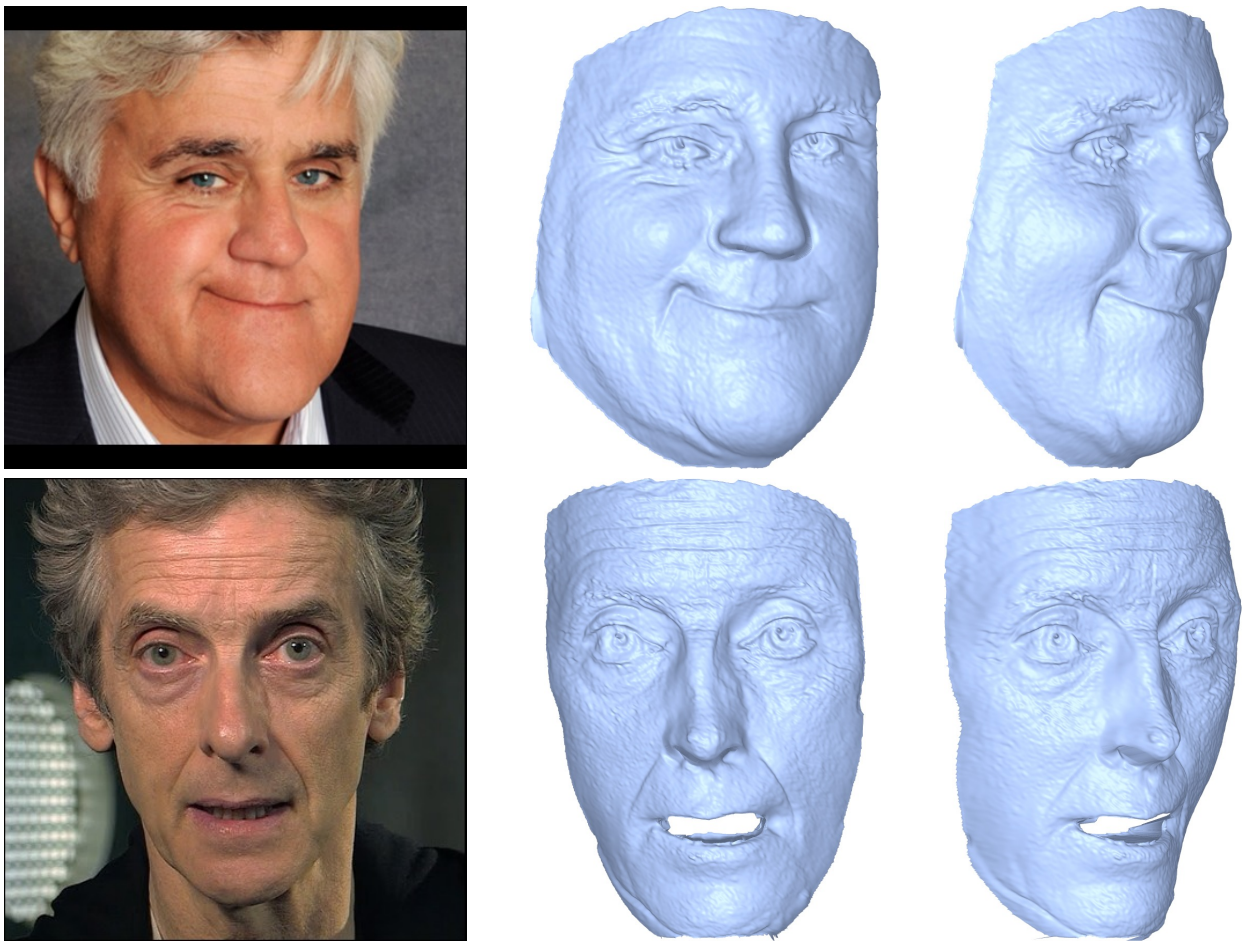

Figure 3: Additional reconstruction results.

Figure 3: Additional reconstruction results.

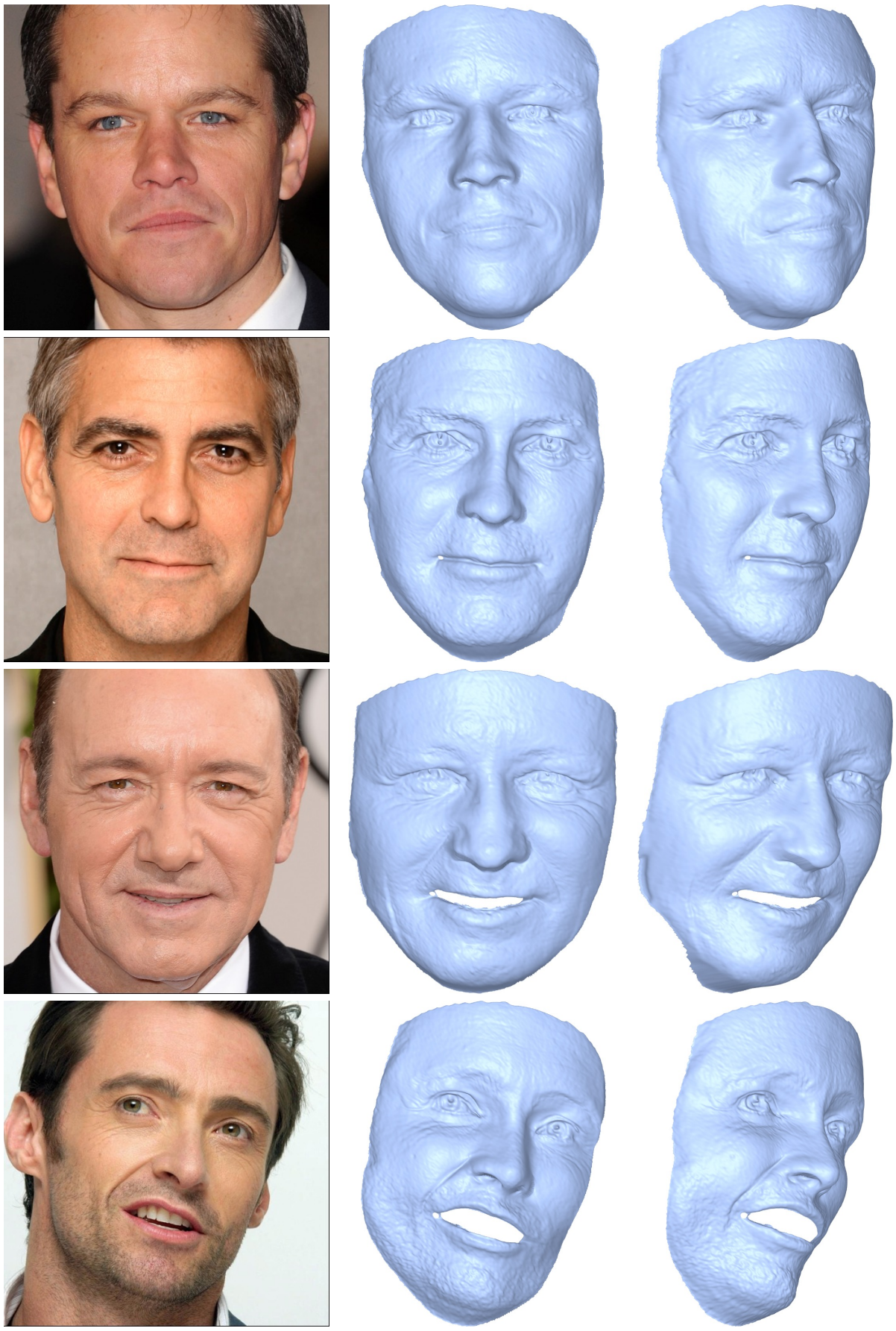

Figure 3: Additional reconstruction results.

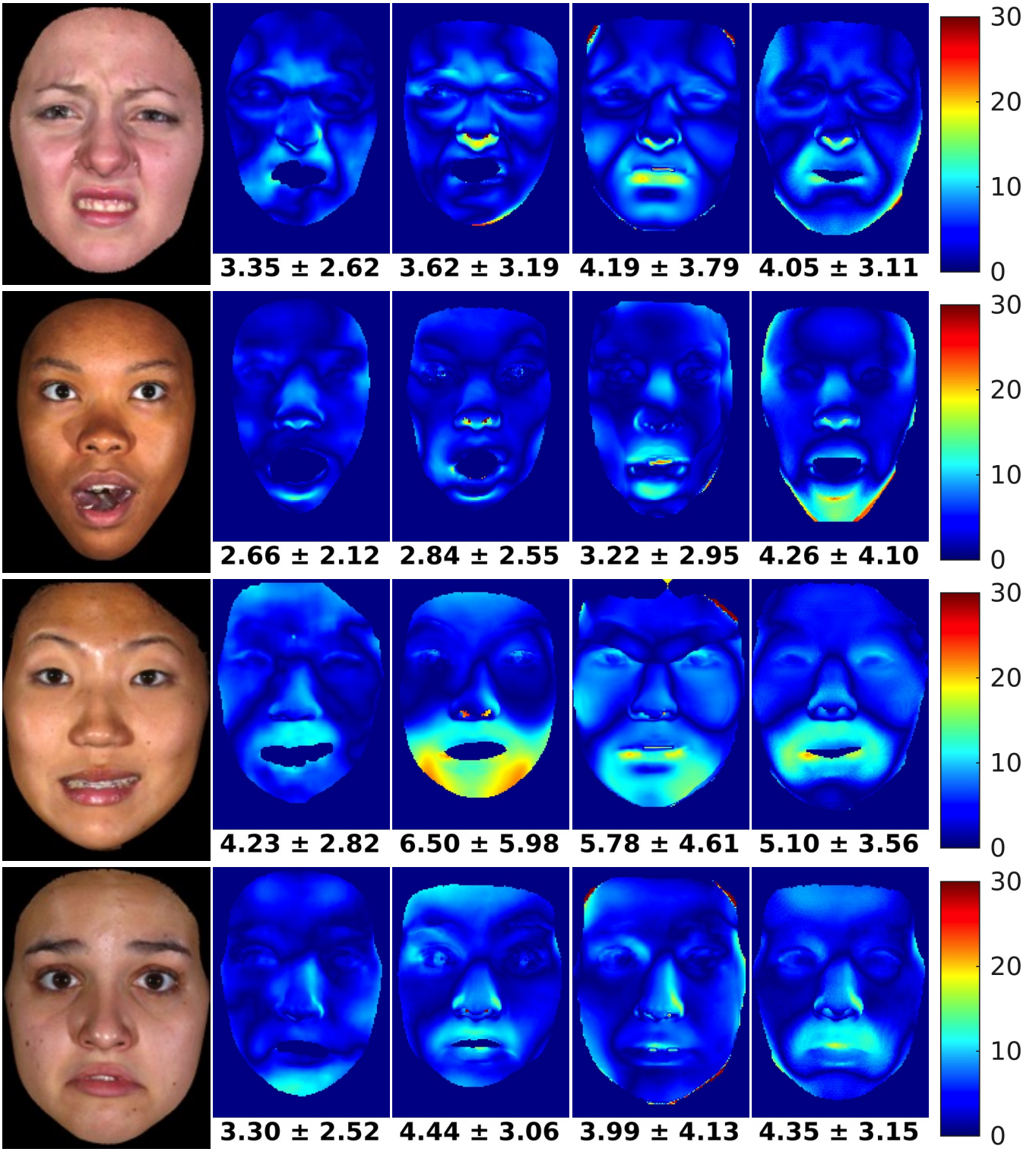

Figure 4: Error heat maps in percentile of ground truth depth.

References

[1] T. Beeler, B. Bickel, P. Beardsley, B. Sumner, and M. Gross. High-quality single-shot capture of facial geometry. In *ACM SIGGRAPH 2010 Papers*, SIGGRAPH '10, pages 40:1–40:9, New York, NY, USA, 2010. ACM.

[2] M. Botsch and O. Sorkine. On linear variational surface deformation methods. *IEEE Transactions on Visualization and Computer Graphics*, 14(1):213–230, Jan 2008.

[3] M. Desbrun, M. Meyer, P. Schröder, and A. H. Barr. Implicit fairing of irregular meshes using diffusion and curvature flow. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '99, pages 317–324, New York, NY, USA, 1999. ACM Press/Addison-Wesley Publishing Co.

[4] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[5] B. T. Helenbrook. Mesh deformation using the biharmonic operator. *International journal for numerical methods in engineering*, 56(7):1007–1021, 2003.

[6] 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.

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

[8] H. Li. *Animation Reconstruction of Deformable Surfaces*. PhD thesis, ETH Zurich, November 2010.

[9] M. Meyer, M. Desbrun, P. Schröder, A. H. Barr, et al. Discrete differential-geometry operators for triangulated 2-manifolds. *Visualization and mathematics*, 3(2):52–58, 2002.

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

[11] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li. High-fidelity pose and expression normalization for face recognition in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 787–796, 2015.