Supplemental Materials:
Topologically Consistent Multi-View Face Inference Using Volumetric Sampling

Tianye Li[1,2], Shichen Liu[1,2], Timo Bolkart[3], Jiayi Liu[1,2], Hao Li[1,2], and Yajie Zhao[1]

[1]USC Institute for Creative Technologies, [2]USC, [3]MPI for Intelligent Systems, Tübingen

Figure 1: Base mesh reconstruction for a multi-view video sequence overlaid on the video frames. Our method captures the facial performance well. The result meshes are temporally stable and accurately align with the input images. Visualizing with a shared checkerboard texture indicates good tracking quality. Please see the *supplemental video* for better visualization.

Supplemental Video. Due to upload size limit, we cannot provide the supplemental video in the official supplemental materials. Please see the video on the project page: https://tianyeli.github.io/tofu.

Additional Quantitative Results. Tab. 1 provides additional quantitative comparisons to other learning based methods, namely 3DMM regression and DFNRMVS [1]. Fig. 2 shows the cumulative error curves for scan-to-mesh distances among the methods. All methods are evaluated on a common held-out test set with 499 ground truth 3D scans; no data of test subjects are used during training. The geometric reconstruction accuracy is evaluated using scan-to-mesh distance (s2m) that measures the distance between each vertex of a ground truth scan, and the closest point in the surface of the reconstructed mesh. The correspondence accuracy is evaluated using a vertex-to-vertex distance (v2v) that measures the distance between each vertex of a registered ground truth mesh, and the semantically corresponding point in the reconstructed mesh.

Methods	median s2m	median v2v
3DMM Regr.	2.104	3.662
3DMM Regr. (PP)	1.659	2.890
DFNRMVS [1] (PP)	1.885	4.565
Our Method	**0.585**	**1.973**

Table 1: Comparison on geometry accuracy (median s2m), correspondence accuracy (median v2v) among the learning based methods, measured in millimeters. "PP" denotes the result after a post-processing Procrustes alignment that solves for the optimal rigid pose (i.e. 3D rotation and translation) and scale to best align the reconstructed mesh with the ground truth. Note that our method requires no post-processing.

Our method outperforms (w/o post-processing) the existing methods (w/ and w/o post-processing) in terms of geometric reconstruction quality and the quality of the correspondence. Note that while the distance of DFNRMVS [1] is higher than for the 3DMM regression, DFNRMVS [1] is visually better in most regions. Their reconstructed meshes tend to have large errors in the forehead and in the jaw areas, as shown in Fig. 5, due to a different mask definition for their on-the-fly deep photo-metric refinement. Fig. 5 in the paper shows that our methods produces significantly better reconstructions than DFNRMVS [1] across the entire face.

Additional Qualitative Results. We evaluate our trained model on a multi-view video sequence with 8 calibrated and synchronized views, captured at 30 fps. We apply our progressive mesh generation network in a frame-by-frame manner, without applying any temporal smoothing. Fig. 1 shows that our base mesh well captures the extreme expressions, and it aligns well with the input images. Despite being trained on static images only, the resulting reconstruction is temporally stable, as shown in the supplemental video. Fig. 9 shows additional base mesh reconstructions for different static multi-view images of varying subjects in different expressions. Our method reconstructs the face shape and expression well, closely to the ground truth scans. We show more visualizations in the *supplemental video*.

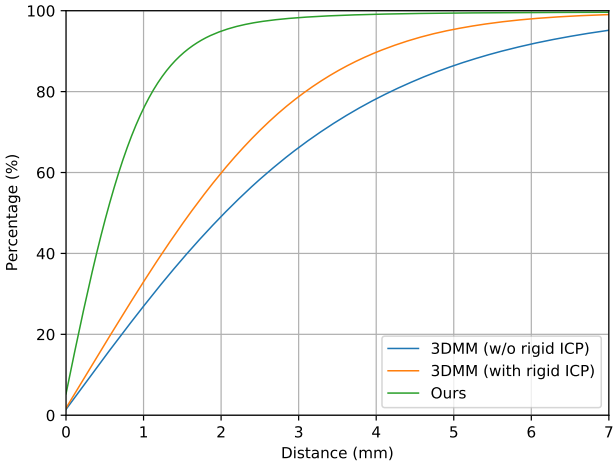

Figure 2: Quantitative evaluation by cumulative error curves for scan-to-mesh distances among learning based methods.

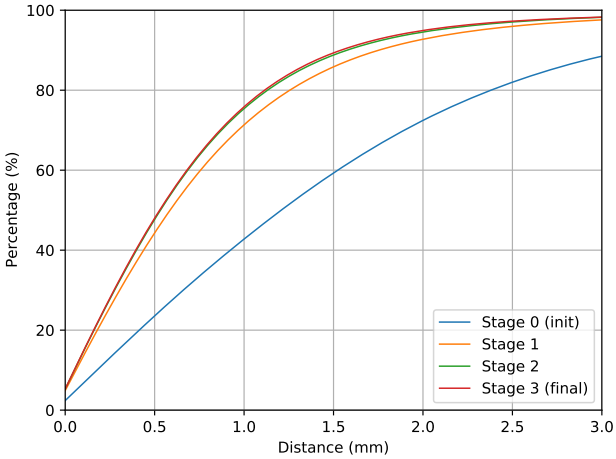

Figure 3: Quantitative evaluation by cumulative error curves for scan-to-mesh distances among local refinement stages.

Impact of Local Refinements. Fig. 3 shows the cumulative error curves for scan-to-mesh distances among the local stages. Given the coarse mesh $\mathcal{M}_0$ as output of the global stage, each local stage successively increases the mesh resolution and refines the vertex locations. Fig. 6 demonstrates the effect of each local refinement step. As shown in Fig. 6, the quality of the reconstructed mesh improves after each local stage, while the scan-to-mesh distance to the scan re-

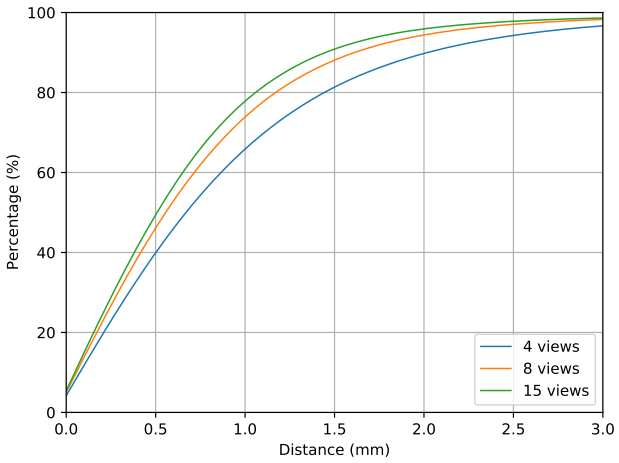

Figure 4: Quantitative evaluation by cumulative error curves for scan-to-mesh distances among various numbers of views.

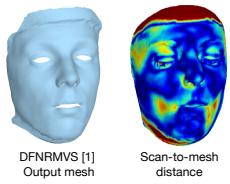

Figure 5: Example results from DFNRMVS [1].

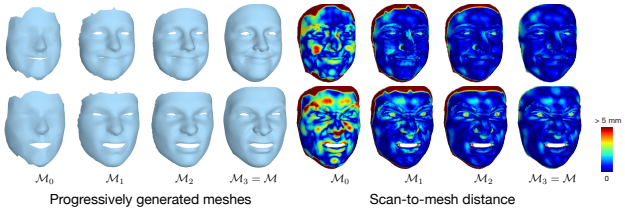

Figure 6: Inferred meshes for global stage $\mathcal{M}_0$ and after upsampling and refinement for each local stage $\mathcal{M}_i$ ($1 \leq i \leq 3$).

duces. Note that details such as nose corners and lips gradually improve through the local stages.

More Ablation on Number of Views. Fig. 4 shows the cumulative error curves for scan-to-mesh distances for networks with different number of input views.

More Results on Appearance and Detail Capture. Fig. 10 shows additional results of the appearance enhancement network, which predicts normal displacements and additional albedo and specular maps on top of the predicted base mesh $\mathcal{M}$ (see Fig. 2 of the paper). Our reconstruction pipeline (i.e. base mesh reconstruction and appearance and detail capture) enables us to reconstruct a 3D face with high-quality assets, 2 to 3 orders of magnitude faster than existing methods, which can readily be used for photorealistic rendering.

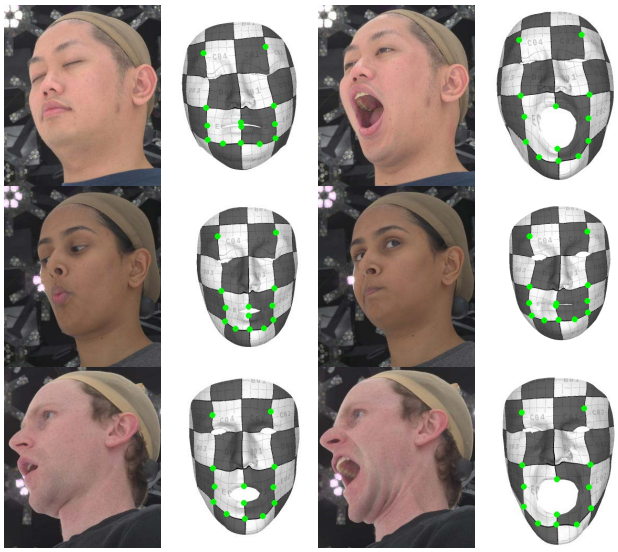

Figure 7: Visualization of cross-subject dense correspondence of the base meshes inferred by ToFu in a shared checkerboard texture.

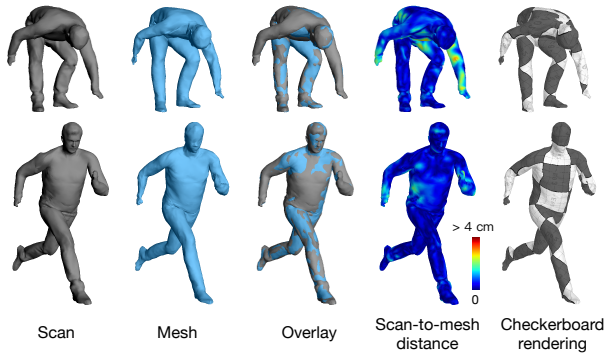

Figure 8: Our system can also infer clothed human body surfaces in consistent topology.

Results on Clothed Human Body Datasets. While we focus on face mesh in correspondence, we find that our method can also predict clothed full body meshes in correspondence. We test our method on a dataset of human bodies as shown in Fig. 8. Human bodies are challenging due to large pose variations and occlusions. Given the challenging inputs, our methods still outputs detailed geometry which closely fit the ground truth surfaces with small scan-to-mesh distances, shown in Fig. 8. Checkerboard projection also shows the accuracy of semantic correspondence among extreme poses. The results demonstrate the flexibility of our method for highly articulated and diverse surfaces.

Albedo. While the input images in our datasets are diffuse albedo images, obtained with polarized lighting and cameras [3, 5], the results, shown in the paper, indicate that our

system can be adapted to non-lightstage setups, e.g. capture system of CoMA [6]. The appearance capture network learns the mapping between albedo images and the details of specular reflectance and fine geometry, as "image-to-image translation". This synthesis is reasonable since the input images contain pore-level details and the outputs are pixel-aligned. However, imperfect albedo images can potentially contain more information on specularity, which in principle can guide the systhesis network to better recover details. This is an interesting perspective and we will explore it as future work.

The $\mathbb{E}$ operator. Let B be batch size and N be vertex number. Given a feature volume $\mathbf{L}_g$ from the global volumetric feature sampling, the global geometry network (3D ConvNet) predicts a probability volume $\mathbf{C}_g$ of size $(B, N, 32, 32, 32)$, whose N-channel is ordered in a predefined vertex order. Finally the soft arg-max operator $\mathbb{E}$ computes the expectations on $\mathbf{C}_g$ *per channel*, and outputs vertices of shape $(B, N, 3)$ corresponding to the predefined order.

On Dense Correspondence. Dense correspondence across identities and expressions is a challenging task [2, 4]. Cross-identity dense correspondence is fundamentally difficult to define beyond significant landmarks, especially in texture-less regions. The state-of-the-art methods rely on landmarks and propagate the dense correspondence by statistical (3DMM) or physical constraints (Laplacian regularization) in a carefully designed optimization process with manual adjustments. Cross-expression correspondence, however definable, can be enforced by photometric consistency (optical flow or differentiable rendering). Our ground truth datasets utilized all these state-of-the-art strategies and therefore can be regarded as one of the best curated datasets. With the "best" ground truth one can get as now, we trained our network in a supervised manner to the ground truth meshes (same topology) with equal vertex weights. Measuring the distances to the ground truth (v2v and landmark errors) gives informative and reliable *cross-expression* evaluations on dense correspondence quality. Furthermore, photometric error visualizations on a shared UV map (as in the main paper) and the stable rendering of reconstructed sequence as in Fig. 1 both qualitatively shows high quality of cross-expression correspondence.

However, quantitative evaluating *cross-identity* dense correspondence is by nature difficult. These two metrics above indirectly measure for cross-subject correspondence. Here we show additional visualizations by rendering inferred meshes in a shared checkerboard texture and highlighting some facial landmarks in Fig. 7. The meshes inferred by ToFu preserve dense semantic correspondences across subjects and expressions, as shown by the landmarks and the uniquely textured regions.

Implementation Details. The appearance enhancement synthesis network uses as similar architecture and losses as proposed by Wang et al. [7]. We train the global generator and 2 multi-scale discriminators at resolution of 512×512. The main difference is that we extract features from two inputs separately before concatenating them and feeding into the convolutional back-end so that we can better encode useful features correspondingly. The network is trained using an Adam optimizer with learning rate of $2e-4$ (decayed from 100 epoch) and batch size of 32 on a NVIDIA GeForce GTX 1080 GPU. For further enhancement, we trained a separate super-resolution network, upsampling attribute maps from 512 to 4K resolution. We modify the network design from ESRGAN [8] by expanding the number of Residual in Residual Dense Blocks (RRDB) from 23 to 32, enabling the upsampling capacity from $4\times$ to $8\times$ in a single pass. The super-resolution network is trained with learning rate of $1e-4$ (halved at 50K, 100K, 200K iterations) and batch size of 16 on two NVIDIA GeForce GTX 1080 GPUs.

References

[1] Ziqian Bai, Zhaopeng Cui, Jamal Ahmed Rahim, Xiaoming Liu, and Ping Tan. Deep facial non-rigid multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5850–5860, 2020. 2, 3

[2] Federica Bogo, Javier Romero, Gerard Pons-Moll, and Michael J Black. Dynamic faust: Registering human bodies in motion. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6233–6242, 2017. 4

[3] Abhijeet Ghosh, Graham Fyffe, Borom Tunwattanapong, Jay Busch, Xueming Yu, and Paul Debevec. Multiview face capture using polarized spherical gradient illumination. *ACM Trans. Graph.*, 30(6), 2011. 3

[4] Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics*, 36(6), 2017. 4

[5] Wan-Chun Ma, Tim Hawkins, Pieter Peers, Charles-Felix Chabert, Malte Weiss, Paul E. Debevec, et al. Rapid acquisition of specular and diffuse normal maps from polarized spherical gradient illumination. *Rendering Techniques*, 2007(9):10, 2007. 3

[6] Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J. Black. Generating 3D faces using convolutional mesh autoencoders. In *Proc. European Conference on Computer Vision (ECCV)*, pages 725–741, 2018. 4

[7] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 4

[8] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *The European Conference on Computer Vision Workshops (ECCVW)*, September 2018. 4

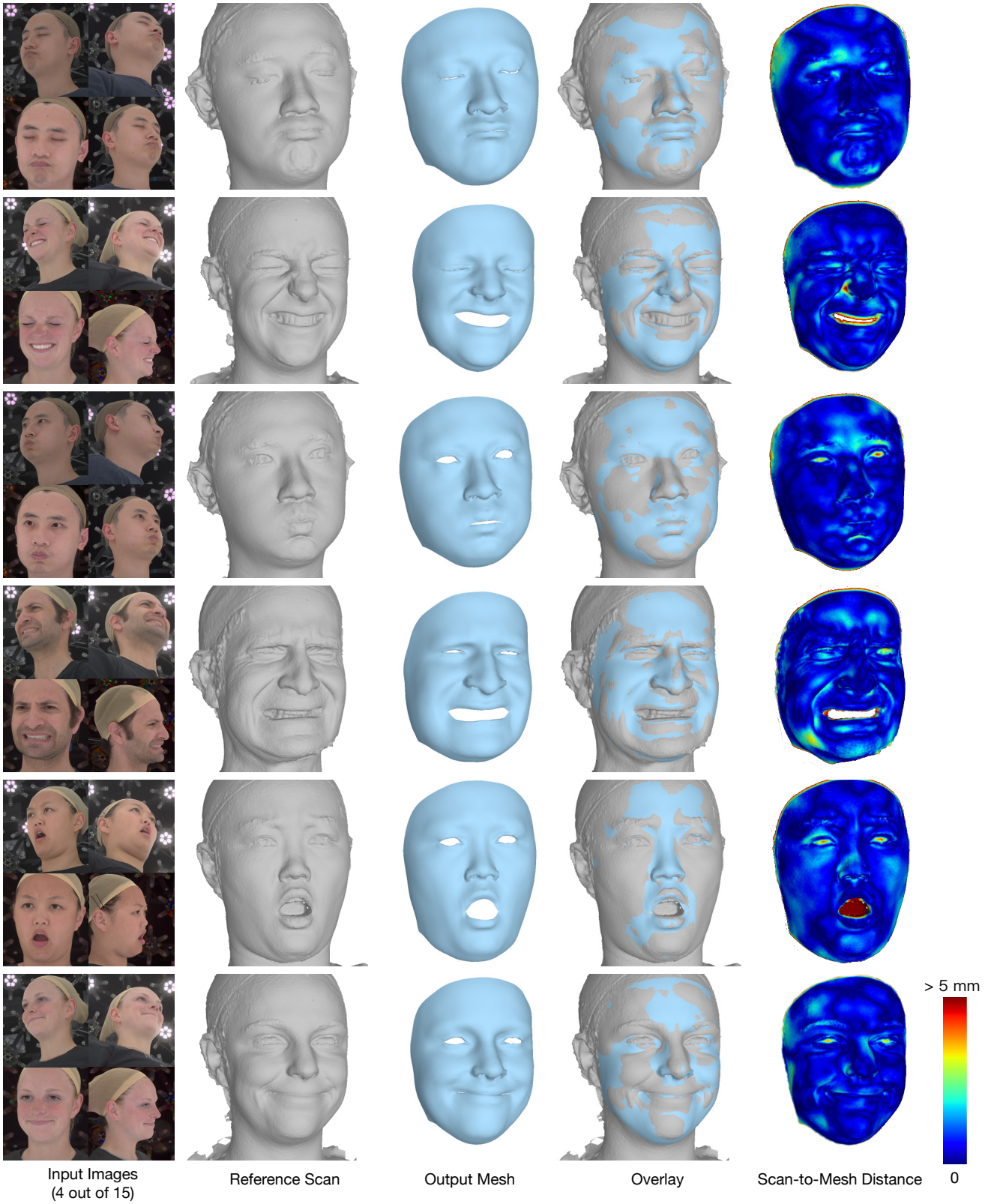

Figure 9: More results of reconstructed meshes in dense correspondence. The scan-to-mesh distance is visualized color coded on the reference scan, where red denotes an error above 5 millimeters.

Figure 10: Our method can generate reliable base alignment meshes, on top of which a comprehensive face modeling pipeline can be built. Here we show more rendering with inferred normal displacements and additional albedo and specular maps.