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: Given (a) multi-view images, our face modeling framework ToFu uses volumetric sampling to predict (b) accurate base meshes in consistent topology as well as (c) high-resolution details and appearances. Our efficient pipeline enables (d) rapid creation of production-quality avatars for animation.

Abstract

High-fidelity face digitization solutions often combine multi-view stereo (MVS) techniques for 3D reconstruction and a non-rigid registration step to establish dense correspondence across identities and expressions. A common problem is the need for manual clean-up after the MVS step, as 3D scans are typically affected by noise and outliers and contain hairy surface regions that need to be cleaned up by artists. Furthermore, mesh registration tends to fail for extreme facial expressions. Most learning-based methods use an underlying 3D morphable model (3DMM) to ensure robustness, but this limits the output accuracy for extreme facial expressions. In addition, the global bottleneck of regression architectures cannot produce meshes that tightly fit the ground truth surfaces. We propose ToFu, ***To****pologically consistent* ***F****ace from m****u****lti-view, a geometry inference framework that can produce topologically consistent meshes across facial identities and expressions using a volumetric representation instead of an explicit underlying 3DMM. Our novel progressive mesh generation network embeds the topological structure of the face in a feature volume, sampled from geometry-aware local features. A coarse-to-fine architecture facilitates dense and accurate facial mesh predictions in a consistent mesh topology. ToFu further captures displacement maps for pore-level geometric details and facilitates high-quality rendering in the form of albedo and specular reflectance maps. These high-quality assets are readily usable by production studios for avatar creation, animation and physically-based skin rendering. We demonstrate state-of-the-art geometric and correspondence accuracy, while only taking 0.385 seconds to compute a mesh with 10K vertices, which is three orders of magnitude faster than traditional techniques. The code and the model are available for research purposes at* https://tianyeli.github.io/tofu.

1. Introduction

Creating high-fidelity digital humans is not only highly sought after in the film and gaming industry, but is also gaining interest in consumer applications, ranging from telepresence in AR/VR to virtual fashion models and virtual assistants. While fully automated single-view avatar digitization solutions exist [29, 30, 43, 57, 64], professional studios still opt for high resolution multi-view images as input, to ensure the highest possible fidelity and surface coverage in a controlled setting [8, 24, 26, 41, 42, 47, 51] instead of unconstrained input data. Typically, high-resolution geometric details ($< 1mm$ error) are desired along with high resolution physically-based material properties (at least $4K$). Furthermore, to build a fully rigged face model for animation, a

large number of facial scans and alignments (often over 30) are performed, typically following some conventions based on the Facial Action Coding System (FACS).

A typical approach used in production consists of using a multi-view stereo acquisition process to capture detailed 3D scans of each facial expression, and a non-rigid registration [8, 37] or inference method [38] is used to warp a 3D face model to each scan in order to ensure consistent mesh topology. Between these two steps, manual clean-up is often necessary to remove artifacts and unwanted surface regions, especially those with facial hair (beards, eyebrows) as well as teeth and neck regions. The registration process is often assisted with manual labeling tasks for correspondences and parameter tweaking to ensure accurate fitting. In a production setting, a completed rig of a person can easily take up to a week to finalize.

Several recent techniques have been introduced to automate this process by fitting a 3D model directly to a calibrated set of input images. The multi-view stereo face modeling method of [22] is not only particularly slow, but relies on dynamic sequences and carefully tuned parameters for each subject to ensure consistent parameterization between expressions. In particular facial expressions that are not captured continuously cannot ensure accurate topological consistencies. More recent deep learning approaches [4, 64] use a 3D morphable model (3DMM) inference to obtain a coarse initial facial expression, but require a refinement step based on optimization to improve fitting accuracy. Those methods are limited in fitting extreme expressions due to the constraints of linear 3DMMs and fitting tightly to the ground-truth face surfaces due to the global nature of their regression architectures. The additional photometric refinement also tends to fit unwanted regions like facial hair.

We introduce a new volumetric approach for consistent 3D face mesh inference using multi-view images. Instead of relying explicitly on a mesh-based face model such as 3DMM, our volumetric approach is more general, allowing it to capture a wider range of expressions and subtle deformation details on the face. Our method is also three orders of magnitude faster than conventional methods, taking only 0.385 seconds to generate a dense 3D mesh (10K vertices) as well as produce additional assets for high-fidelity production use cases, such as albedo, specular, and high-resolution displacement maps.

To this end, we propose a progressive mesh generation network that can infer a topologically consistent mesh directly. Our volumetric architecture predicts vertex locations as probability distributions, along with volumetric features that are extracted using the underlying multi-view geometry. The topological structure of the face is embedded into this architecture using a hierarchical mesh representation and coarse-to-fine network.

Our experiments show that ToFu is capable of producing highly accurate geometry consistent with topology automatically, while existing methods either rely on manual clean-up and parameter tuning, or are less accurate especially for subjects with facial hair. Since we can ensure a consistent parameterization across facial identities and expressions without any human input, our solution is suitable for scaled digitization of high-fidelity facial avatars, We not only reduce the turn around time for production, but is also provide a critical solution for generating large facial datasets, which is often associated with excessive manual labor. Our main contributions are:

- A novel volumetric feature sampling and refinement model for topologically consistent 3D mesh reconstruction from multi-view images.
- An appearance capture network to infer high-resolution skin details and appearance maps, which, combined with the base mesh, forms a complete package suitable for production in animation and photorealisitic rendering.
- We demonstrate state-of-the-art performance for combined geometry and correspondence accuracy, while achieving mesh inference at near interactive rates.
- Code and model are publicly available.

2. Related Work

Face Capture. Traditionally, face acquisition is separated into two steps, 3D face reconstruction and registration [18]. Facial geometry can be captured with laser scanners [36], passive Multi-View Stereo (MVS) capture systems [7], dedicated active photometric stereo systems [24, 42], or depth sensors based on structured light or time-of-flight sensors. Among these, MVS is the most commonly used [19, 21, 25, 35, 44, 61]. Although these approaches produce high-quality geometry, they suffer from heavy computation due to the pairwise features matching across views, and they tend to fail in case of sparse view inputs due to the lack of overlapping neighboring views. More recently, deep neural networks learn multi-view feature matching for 3D geometry reconstruction [27, 32, 34, 52, 65]. Compared to classical MVS methods, these learning based methods represent a trade-off between accuracy and efficacy. All these MVS methods output unstructured meshes, while our method produces meshes in dense vertex correspondence.

Most registration methods use a template mesh and fit it to the scan surface by minimizing the distance between the scan's surface and the template. For optimization, the template mesh is commonly parameterized with a statistical shape space [3, 9, 11, 39] or a general blendshape basis [49]. Other approaches directly optimize the vertices of the template mesh using a non-rigid Iterative Closest Point (ICP)

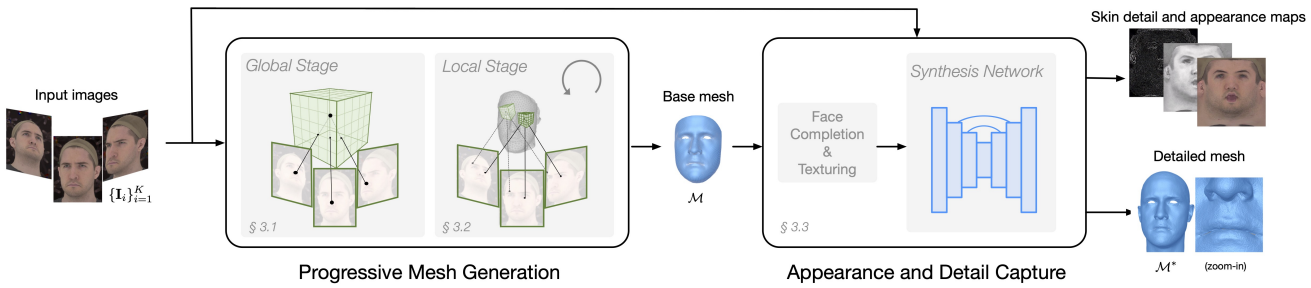

Figure 2: Overview of our end-to-end face modeling system. Given images captured from multi-views, the progressive mesh generation network predicts an accurate face mesh in consistent topology. Then the appearance and detail capture network synthesizes high-resolution skin detail and attribute maps, which enables highly detailed geometry and photo-realistic renderings.

[37], with a statistical model as regularizer [40], or jointly optimize correspondence across an entire dataset in a group-wise fashion [13, 66]. For a more thorough review of face acquisition and registration, see Egger et al. [18]. All these registration methods solve for facial correspondence independent from the data acquisition. Therefore, errors in the raw scan data propagate into the registration.

Only few methods exist that are similar to our method of directly outputting high-quality registered 3D faces from calibrated multi-view input [8, 14, 15, 22]. While sharing a similar goal, our method goes beyond these approaches in several significant ways. Unlike our method, they require calibrated multi-view image sequence input, contain multiple optimization steps (e.g. for building a subject specific template [22], or anchor frame meshes [8]), and are computationally slow (e.g. 25 minutes per frame for the coarse mesh reconstruction [22]). ToFu instead takes calibrated multi-view images as input (i.e. static) and directly outputs a high-quality mesh in dense vertex correspondence in 0.385 seconds. Regardless, our method achieves stable reconstruction and registration results for sequence input.

Model-based reconstruction. A large body of work aims at reconstructing 3D faces from unconstrained images or monocular videos. To constrain the problem, most methods estimate the coefficients of a statistical 3D morphable models (3DMM) in an optimization-based [1, 6, 10, 11, 58] or learning-based framework [16, 20, 23, 46, 50, 57, 59]. Due to the use of over-simplified, mostly linear statistical models, the reconstructed meshes only capture the coarse geometry shape while subtle details are missing. For better generalization to unconstrained conditions, [54, 60] jointly learn a 3D prior and reconstruct 3D faces from images. Although monocular reconstruction methods can provide visually appealing 3D face reconstructions, their accuracy and quality is not suitable for applications which require metrically accurate geometry. Recently published work indicates that existing state-of-the-art monocular 3D face reconstructions are metrically worse or only marginally better compared to a static model mean face, when compared to ground truth 3D scans [50]. This comes at little surprise as inferring 3D geometry from a single image is an ill-posed problem due to the inherent ambiguity of focal length, scale and shape [5] as under perspective projection different shapes result in the same image for different object-camera distances. Our method instead leverages explicit calibrated multi-view information to reconstruct metrically accurate 3D geometry.

3. Multi-View Face Inference

As shown in Fig. 2, given images $\{\mathbf{I}_i\}_{i=1}^K$ in K views with known camera calibration $\{\mathbf{P}_i\}_{i=1}^K$, together denoted as $\mathcal{I} = \{\mathbf{I}_i, \mathbf{P}_i\}_{i=1}^K$, the goal of ToFu is two-fold: (1) to reconstruct an accurate *base mesh* in an artist-designed topology, and (2) to estimate pore-level geometric details and high-quality facial appearance in form of albedo and specular reflectance maps. Formally, an output base mesh $\mathcal{M}$ contains a list of vertices $\mathbf{V} \in \mathbb{R}^{N\times 3}$ and a fixed triangulation $\mathbf{T}$. The base meshes are required to (1) tightly fit the face surfaces, (2) share a common artist-designed mesh topology, where each vertex encodes the same semantic interpretation across all meshes, and (3) have a sufficient triangle or quad density (with $N > 10^4$ number of vertices).

The key to dense mesh prediction is a coarse-to-fine network architecture, as shown in Fig. 3. The desired semantic mesh correspondence is naturally embedded in the hierarchical architecture. Based on that, the geometry is inferred by the following two stages: (1) a coarse mesh prediction $\mathcal{M}_0$, by the global stage $\mathbf{V}_0 = \mathcal{F}_g(\mathcal{I})$; and (2) iteratively upsampling and refining into the denser meshes $\{\mathcal{M}_1, \mathcal{M}_2, ..., \mathcal{M}_L\}$, by the local stage $\mathbf{V}_{k+1} = \mathcal{F}_l(\mathcal{I}, \mathbf{V}_k)$. $\mathcal{M}_L$ is the final prediction of base mesh $\mathcal{M}$.

Conceptually, the global stage mimics a learning-based MVS, while the local stage provides "updates" as if in an iterative mesh registration. In contrast to the two traditional methods, our two steps share consistent correspondence in a fixed topology and use volumetric features for geometry inference and surface refinement.

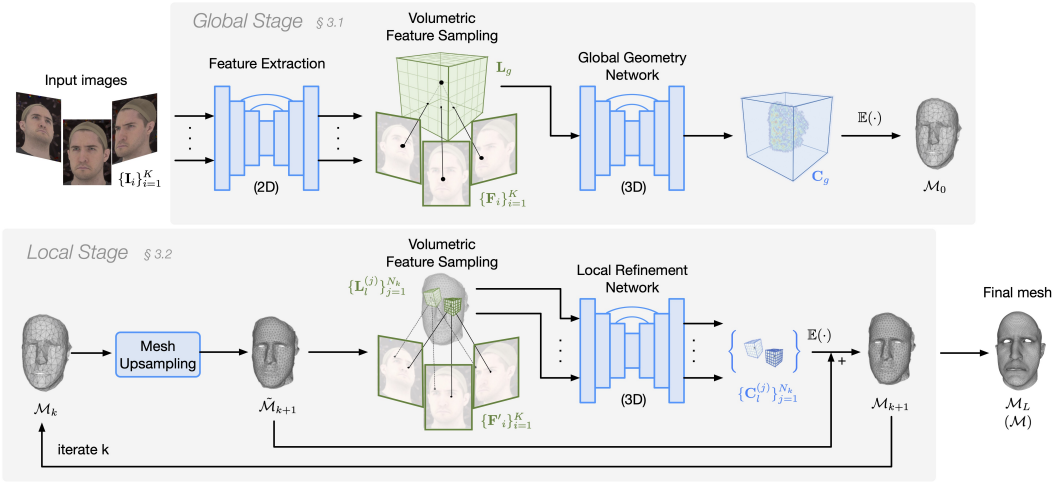

Figure 3: Overview of the progressive mesh generation network.

3.1. Global Geometry Stage

Volumetric Feature Sampling. In order to extract salient features to predict surface points in correspondence, we deploy a shared U-Net convolutional network to extract local 2D feature maps $\mathbf{F}_i$ for each input image $\mathbf{I}_i$. We sample volumetric features $\mathbf{L}$ by bilinearly sampling and fusing image features at projected coordinates in all images for each local point $\mathbf{v} \in \mathbb{R}^3$ in the 3D grid $\mathcal{G}$:

$$\mathbf{L}(\mathbf{v}) = \sigma(\{\mathbf{F}_i(\Pi(\mathbf{v}, \mathbf{P}_i))\}_{i=1}^{K}), \tag{1}$$

where $\Pi(\cdot)$ is the perspective projection function and $\sigma(\cdot)$ is a view-wise fusion function, for which common choices can be max, mean or standard deviation. The 3D grid $\mathcal{G}$ is a set of points on a regular 3D grid, which can be defined at arbitrary locations with arbitrary shapes. Here we choose cube grids, as shown in green cubes in Fig. 3 to feed into 3D convolution networks.

Global Geometry Network. To enable the vertex flexibility, we design the network to predict vertex location free of the constraint of 3DMMs. To encourage better generalization, we design a volumetric network architecture to learn the probabilistic distribution instead of the absolute location for each vertex. We define a canonical global grid $\mathcal{G}_g$ that covers the whole captured volume for subject heads. We apply the volumetric feature sampling (Eq. 1) on the global grid $\mathcal{G}_g$ to obtain the global volumetric feature $\mathbf{L}_g$, similar to [31, 33]. We deploy the global geometry network Φ_g, a 3D convolutional network with skip connections, to predict a probability volume $\mathbf{C}_g = \Phi_g(\mathbf{L}_g)$, in which each channel encodes the probability distribution for the location of a corresponding vertex in the initial mesh $\mathcal{M}_0$. The vertex locations are extracted by a per-channel soft-argmax operation, $\mathbf{V}_0 = \mathbb{E}(\mathbf{C}_g)$, similar to that in [33].

3.2. Local Geometry Stage

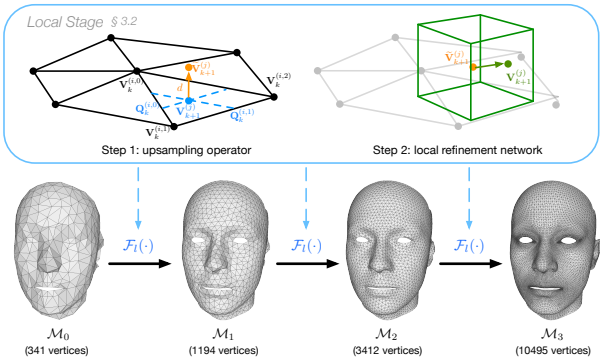

Figure 4: The iterative upsampling and refinement process in the local geometry stage.

Based on the coarse mesh $\mathcal{M}_0$ obtained from the global stage, the local stage progressively produces meshes in higher resolution and with finer details, $\{\mathcal{M}_k\}_{k=1}^{L}$. At each level k, this process is done in two steps, as shown in Fig. 4: (1) a fixed and differentiable upsampling operator to provide a reliable initialization for upsampled meshes, and (2) a local refinement network to further improve the surface details based on the input images.

Upsampling Operator. Ranjan et al. [45] propose a mesh upsampling technique based on the barycentric embedding of vertices in the lower-resolution mesh version. Directly using this upsampling scheme results in unsmooth artifacts, as the barycentric embedding contrains the upsampled vertices to lie in the surface of the lower-resolution mesh. Instead, we use additional normal displacement weights as shown in step 1 of Fig. 4. Given a sparser mesh $\mathcal{M}_k = (\mathbf{V}_k, \mathbf{T}_k)$ and its per-vertex normal vectors $\mathbf{N}_k$, we upsam-

ple the mesh by

$$\widetilde{\mathbf{V}}_{k+1} = \mathbf{Q}_k \mathbf{V}_k + \mathbf{D}_k \mathbf{N}_k, \tag{2}$$

where $\mathbf{Q}_k \in \mathbb{R}^{N_{k+1} \times N_k}$ is the barycentric weight matrix as in [45] and $\mathbf{D}_{k-1} \in \mathbb{R}^{N_{k+1} \times N_k}$ is the additional coefficient matrix that apply displacement vectors along normal directions. The normal displacements encode additional surface details that allow vertices to be outside of the input surface.

For a hierarchy with L levels, we first downsample the full-resolution template mesh $\mathcal{T} = (\mathbf{V}, \mathbf{T}) := \mathcal{T}_L$ by isotropic remeshing and non-rigid registration, into a series of meshes with decreasing resolution while still preserving geometry and topology of original mesh: $\{\mathcal{T}_{L-1}, \mathcal{T}_{L-2}, ..., \mathcal{T}_0\}$. Next, we embed the vertices at higher resolution in the surface at lower resolution meshes by barycentric coordinates $\mathbf{Q}_k$ as in [45]. We then project the remaining residual vectors onto the normal direction and obtain $\mathbf{D}_k$.

Local Refinement Network. Around each vertex (indexed with j) of the upsampled mesh $\widetilde{\mathbf{V}}_{k+1}^{(j)}$, we define a smaller grid than $\mathcal{G}_g$ in the global stage in the local neighborhood $\mathcal{G}_l^{(j)}$. We sample local volumetric features $\mathbf{L}_l^{(j)}$ by Eq. 1. For each local feature volume, we apply the local refinement network Φ_l, a 3D convolutional network with skip connections, to predict per-vertex probability volume $\mathbf{C}_l^{(j)} = \Phi_l(\mathbf{L}_l^{(j)})$. Then we compute the corrective vector by the expectation operator, $\delta\mathbf{V}_{k+1}^{(j)} = \mathbb{E}(\mathbf{C}_l^{(j)})$. This process is applied to all vertices independently, and therefore can be parallelized in batches. Finally the upsampled and refined mesh vertices are

$$\mathbf{V}_{k+1} = \widetilde{\mathbf{V}}_{k+1} + \delta\mathbf{V}_{k+1}. \tag{3}$$

Given $\mathcal{M}_0$, we iteratively apply the local stage at all levels until we reach the highest resolution and obtain $\mathcal{M}_L$.

The volumetric feature sampling and the upsampling operator, along with the networks are fully differentiable, enabling the progressive geometry network end-to-end trainable from input images to dense registered meshes.

3.3. Appearance and Detail Capture

Skin detail and appearance maps are commonly used in photo-realistic rendering, which is often difficult to estimate without special capture hardware, such as the Light Stage capture system [17]. We propose a simple yet effective architecture to estimate high-resolution detail and appearance maps, potentially without the dependency on special appearance capture systems.

Albedo Maps Generation. The base meshes are reconstructed for a smaller head region. We augment the base meshes by additional fitting for the back of the head using Laplacian deformation [53]. We then perform the standard texturing given the completed mesh and multi-view images and obtain the albedo reflectance map on the UV domain. Furthermore, by applying the texturing process and sample vertex locations instead of RGB colors, we obtain another map on the UV domain, that we call the geometry map.

Detail Maps Synthesis. To further augment the representation, we adopt an image-to-image translation strategy to infer finer-level details. Using a network similar to [62], our synthesis network infers specular reflectances and displacements given both albedo and geometry map. We then upscale all the texture maps to 4K resolution by using the super resolution strategy of [63]. We can obtain the detailed mesh in high-resolution by applying the displacement maps on the base mesh, as shown in Fig. 2. The reconstructed skin detail and appearance maps are directly usable for standard graphics pipelines for photo-realistic rendering.

4. Experiments

Datasets. ToFu is trained and evaluated on datasets captured with a Light Stage system [24, 42], with 3D scans from MVS, ground truth base meshes from a traditional mesh registration pipeline [39], and ground truth skin attributes from the traditional light stage pipeline [17]. In particular, we correct the ground truth base meshes with optical flow and manual work of a professional artist, to ensure high quality and high accuracy of registration. The dataset contains 64 subjects (45 for train and 19 for test), covering a wide diversities in gender, age and ethnicity. Each set of capture contains a neutral face and 26 expressions, including some extreme face deformations (e.g. mouth widely open), asymmetrical motions (jaw to left/right) and subtle expressions (e.g. concave cheek or eye motions).

Implementation Details. For the progressive mesh generation network, our feature extraction network adopts a pre-trained UNet [48] with ResNet34 [28] as its backbone, which predicts feature maps of same resolution of input image with 8 channels. The volumetric features of the global stage are sampled from a 32^3 grid with grid size of 10 millimeters, the local stage uses a 8^3 grid with a grid size of 2.5 millimeters. We randomly rotate the grids for the volumetric feature sampling as data augmentation during training. The mesh hierarchy with $L = 3$ contains meshes with 341, 1194, 3412 and 10495 vertices. Both the global geometry network and local refinement network use a similar architecture as the V2V network in [33]. Both stages are trained separately. The global stage trains for 400K iterations with a l_2 loss $\left\|\mathbf{V}_0 - \bar{\mathbf{V}}_0\right\|_2^2$, the local stage trains for 150K iterations with a l_2 loss combined across mesh hierarchy levels with equal weights, $\sum_{k=0}^{L} \left\|\mathbf{V}_k - \bar{\mathbf{V}}_k\right\|_2^2$, where $\bar{\mathbf{V}}_k$ is the ground truth base mesh vertices for the predicted $\mathbf{V}_k$ at level k. We train the progressive mesh generation network using Adam optimizer with a learning rate of $1e-4$ and batch

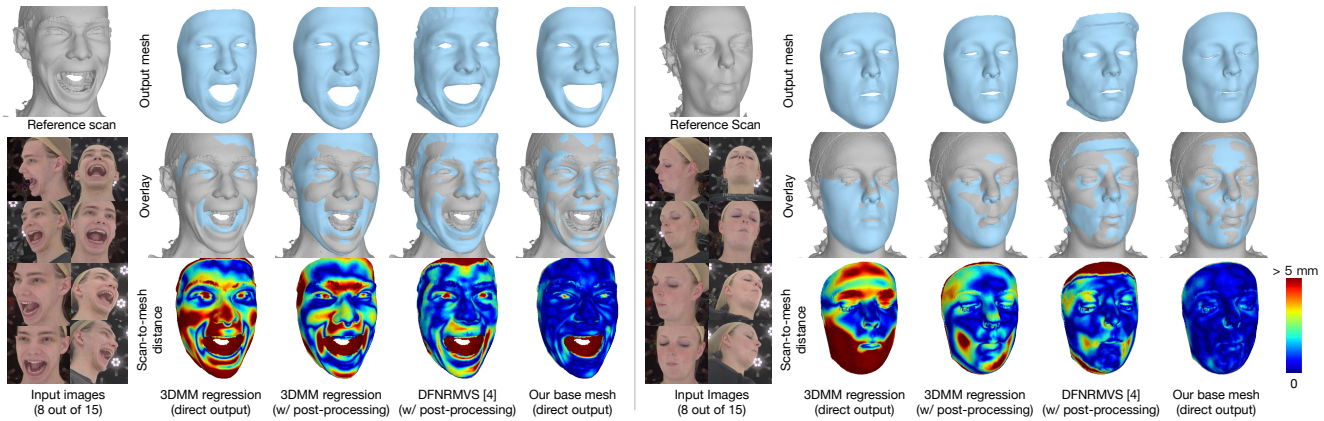

Figure 5: Qualitative comparison on geometric accuracy with the existing methods. The scan-to-mesh distance is visualized in heatmap (red means $>$ 5 mm). Note that 3DMM and DFNRMVS [4] need rigid ICP as post-processing. Our outputs require no post-processing, while outperforming the existing learning-based method in geometry accuracy.

size of 2 on a single NVIDIA V100 GPU. For the detail maps synthesis, we adopt the synthesis network from [62] and the super-resolution network from ESRGAN [63]. For more details, see the *Sup. Mat.*

4.1. Results

Baselines. We evaluate the performance of our base mesh prediction and compare to the following existing methods: (1) **Traditional MVS and Registration:** we run commercial photogrammetry software AliceVision [2], followed by non-rigid ICP surface registration. (2) **3DMM Regression:** we adopt a network architecture similar to [55, 56, 64] for a multi-view setting. (3) **DFNRMVS [4]:** a method that learns an adaptive model space and on-the-fly iterative refinements on top of 3DMM regression.

We argue that the two-step methods of MVS and registration is susceptible to MVS errors and requires manual tweaking optimization parameters for different inputs, which makes it not robust. Our method shows robustness and generalizability for challenging cases, outperforms existing learning-based method and achieves the state-of-the-art geometry and correspondence quality. Our method has efficient run-time. We show various ablation studies to validate the effectiveness of our design. We will provide more comparison and results in the *Sup. Mat.*

Robustness. Fig. 6 show the results from various methods given challenging inputs. Note that when the nose of the subject (top case) is specular reflective (due to oily skin) or has facial hair, the traditional MVS fails to reconstruct the true surface, producing artifacts that affect the subsequent surface registration step. With conservative optimization parameters (e.g. strong reliance on 3DMM), the result is more robust. However with the same parameters, it affects the flexibility for fitting detailed shape and motion for other input cases (e.g. bottom case). Furthermore, the

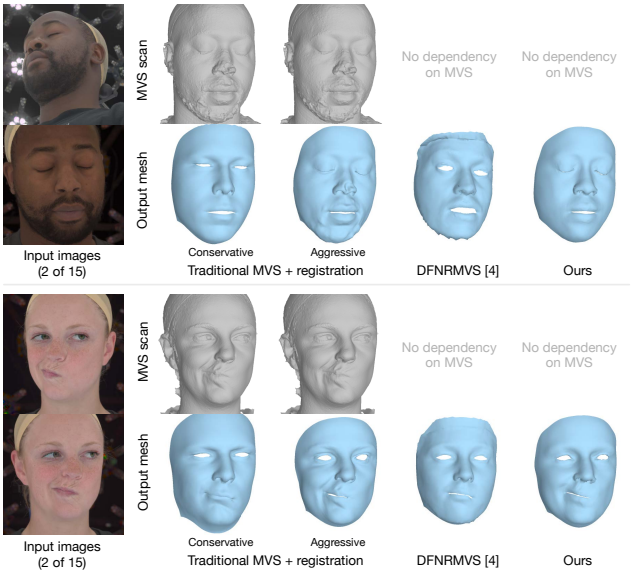

Figure 6: Evaluation on method robustness.

extreme and asymmetrical motion is challenging for fitting only within the morphable model. This case requires "aggressive" fitting, in which less regularizations are applied. Therefore we point out this dilemma of general parameters in the traditional MVS and registration affects of automation and requires much manual work for high-quality results. The learning based method DFNRMVS [4] shows potential for robustness and generalizability. However, they cannot output meshes in accurate shape and expressions. On the contrary, our model shows superior performances in predicting a reliable mesh, given such challenging inputs. Note that the details, such as closed eyelids and asymmetrical mouth motion are faithfully captured.

Geometric Accuracy. Fig. 5 shows the inferred meshes

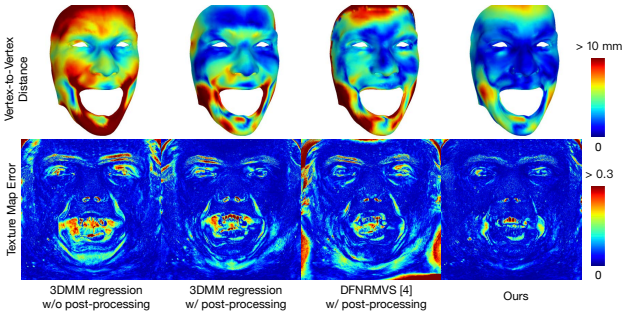

Figure 7: Visualization on correspondence accuracy

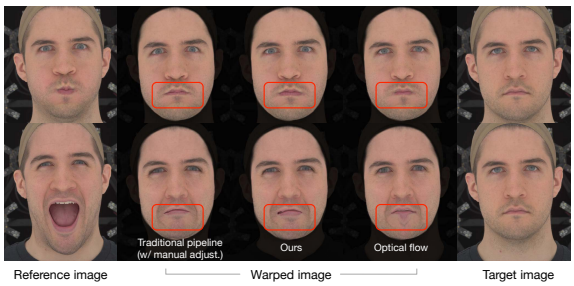

Figure 8: Qualitative evaluation on correspondence compared to optical flow.

Methods	Time	Automatic
Traditional pipeline	600+	✗
DFNRMVS [4]	4.5	✓
ToFu (base mesh)	0.385	✓

Table 1: Comparison on run time on base mesh, given images from 15 views and measured in seconds.

given images from 15 views, along with error visualizations with the reference scans. The 3DMM regression method cannot fit extreme or subtle expressions (wide mouth open, concave cheek and eye shut). The adaptive space and the online refinement improve DFNRMVS [4] for a better fitting, but it still lacks the accuracy to cover the geometric details. Our method is capable of predicting base meshes that closely fit the ground truth surfaces. The results recover identities for the subjects and captures challenging expressions such as extreme mouth opening or subtle non-linearity of small muscles movement (concave cheek) which cannot be modeled by linear 3DMMs. The overlay and error visualizations indicate that our reconstruction fits the ground truth scan closely with fitting errors significantly below 5 millimeters. Due to not being able to utilize true projection parameters, the results of 3DMM regression and DFNRMVS [4] lack accuracy in absolute coordinate and need a Procrustes analysis (scale and rigid pose) as post-processing for further fitting to the target. In contrast, our method outperforms these methods *without* post-processing.

As a quantitative evaluation, we measure the distribution of scan-to-mesh distances. 78.3% of vertices by our methods have scan-to-mesh distance lower than 1 mm. This result outperforms the 3DMM regression which have 27.0% and 33.1% (without and with post-processing). The median scan-to-mesh distance for our results is 0.584 mm, achieving sub-millimeter performance. We show cumulative scan-to-mesh distance curves in the *Sup. Mat.*

Correspondence Accuracy. We provide quantitative measure for correspondence accuracy for generated base meshes, by comparing them to the ground truth aligned meshes (artist-generated in the same topology), and compute the vertex-to-vertex (v2v) distances on a test set. The 3DMM regression method achieves a median v2v distance of 3.66 mm / 2.88 mm (w/o and w/ post-processing). Our method achieves 1.97 mm outperforming the existing method. The v2v distances are also visualized on the ground truth mesh in Fig. 7. We additionally evaluate our aligned meshes by the median errors to the ground truth 3D landmarks. Our method achieves 2.02 mm, while the 3DMM regression method achieves 3.92 mm / 3.21 mm (w/o and w/ post-processing). We provide more quantitative evaluations in the *Sup. Mat.*

We compute the photometric errors between the texture map of the output meshes and the one of the ground truth meshes. Lower photometric errors indicate the UV textures match the pre-designed UV parametrization (i.e. better correspondence). Our method has significantly lower errors, especially in the eyebrow region, around the jaw and for wrinkles around eyes and nose. Note that the 3DMM regression method without post-processing performs worse, while our method requires no post-processing.

In Fig. 8, we further evaluate the correspondence quality by projecting it onto 2D images and warping the reference image (extreme expression) back to target image (neutral expression). The ideal warping outputs would be as close to the target image as possible, except for shades as in wrinkles. We compare the performance with traditional pipeline of MVS and registration (with manual adjustment) and the traditional optical flow method. Our method recovers better 2D correspondence than optical flow, which relies on local matching which tends to fail when occlusion and large motion, as shown in Fig. 8 (see lip regions). Further optical flow takes 30 seconds on image resolution 1366 × 1003, compared within 1 second based on our base meshes. The traditional method achieves good results, but at a cost of 3 orders of magnitude longer of processing time and possibly manual adjustment.

Inference Speed. The traditional pipeline takes more than 10 minutes and potentially more time for manual adjustments. DFNRMVS [4] infers faces without tuning at test-time but is still slower at 4.5 seconds due to its online op-

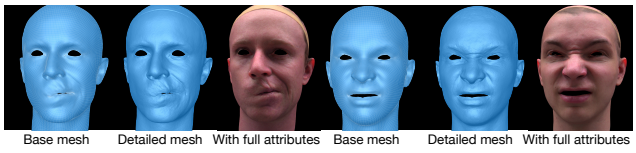

Figure 9: Based on our reliable base meshes, our appearance and detail capture network predicts realistic face skin details and attributes, without special hardware such as Light Stage at test-time, enabling photo-realistic rendering.

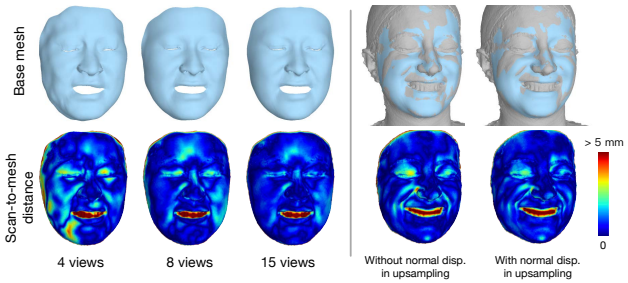

Figure 10: Ablation studies. Left: number of input camera views; Right: on normal displacement weights in mesh upsampling function.

Figure 11: Results on CoMA [45] datasets.

timization step and heavy computation on the dense photometric term. Our global and local stage takes 0.081 seconds and 0.304 seconds respectively. As shown in Table 1, our method produces a high-quality registered base mesh in 0.385 seconds, and achieves sub-second performance, while being fully automatic without manual tweaking.

Appearance Capture. In Fig. 1 and Fig. 9, we show rendering results with the inferred displacement and albedo and specular maps, enabling photo-realistic renderings.

Ablation Studies. In Fig. 10 (left), we evaluate the robustness of our network on various numbers of input views. The resulting quality degrades gracefully as the views decrease. Our method produces reasonable results on views as sparse as 4, which is extremely difficult for standard MVS due to large baseline and little overlaps. Fig. 10 (right) demonstrates the normal displacement in the upsampling function contributes in capturing fine shape details. We provide more ablation studies in the *Sup. Mat.*

Generalization to New Capture Setups. We finetune our network on the CoMA [45] dataset, which contains a different camera setup, significantly fewer views (4) and subjects (12), different lighting conditions and special make-up patterns painted on subjects' faces. The results in Fig. 11 show that our system can in principle be applied to the different capture setups. However, we observe some artifacts around jaws and slightly protruding eyebrow bones. This is potentially due to limited number of subjects and insufficient camera coverage (e.g. the 3rd image misses the jaw region).

5. Conclusion

We introduced a 3D face inference approach from multi-view input images that can produce high-fidelity 3D faces meshes with consistent topology using a volumetric sampling approach. We have shown that, given multi-view inputs, implicitly learning a shape variation and deformation field can produce superior results, compared to methods that use an underlying 3DMM even if they refine the resulting inference with an optimization step. We have demonstrated sub-millimeter surface reconstruction accuracy, and state-of-the-art correspondence performance while achieving up to 3 orders of magnitude of speed improvement over conventional techniques. Most importantly, our approach is fully automated and eliminates the need for data clean up after MVS, or any parameter tweaking for conventional non-rigid registration techniques. Our experiments also show that the volumetric feature sampling can aggregate effectively features across views at various scales and can also provide salient information for predicting accurate alignment without the need for any manual post-processing. Our next step is to extend our approach to regions beyond the skin region, including teeth, tongue, and eyes. We believe that our volumetric digitization framework can handle non-parametric facial surfaces, which could potentially eliminate the need for specialized shaders and models in conventional graphics pipelines. Furthermore, we would like to explore video sequences, and investigate ways to ensure temporal coherency in fine-scale surface deformations. Our model is suitable for articulated non-rigid objects such as human bodies, which motivates us to look into more general shapes and objects such as clothing and hair.

Acknowledgement. We thank M. Ramos, M. He and J. Yang for the help in visualizations, and P. Prasad, Z. Li and Z. Lv for proofreading. The research was sponsored by the Army Research Office and under Cooperative Agreement Number W911NF-20-2-0053, and sponsored by the U.S. Army Research Laboratory (ARL) under contract number W911NF-14-D-0005, the CONIX Research Center, one of six centers in JUMP, a Semiconductor Research Corporation (SRC) program sponsored by DARPA and in part by the ONR YIP grant N00014-17-S-FO14. Statements and opinions expressed and content included do not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

Disclosure. While TB is a part-time employee of Amazon, his research was performed solely at, and funded by, MPI.

References

[1] Oswald Aldrian and William AP Smith. Inverse rendering of faces on a cloudy day. In *Proc. European Conference on Computer Vision (ECCV)*, pages 201–214, 2012. 3

[2] Alicevision. Alicevision, 2020. 6

[3] Brian Amberg, Reinhard Knothe, and Thomas Vetter. Expression invariant 3D face recognition with a morphable model. In *International Conference on Automatic Face Gesture Recognition*, pages 1–6, 2008. 2

[4] 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, 6, 7, 12, 14

[5] Anil Bas and William A. P. Smith. What does 2d geometric information really tell us about 3d face shape? *International Journal of Computer Vision*, 127, 2019. 3

[6] Anil Bas, William A. P. Smith, Timo Bolkart, and Stefanie Wuhrer. Fitting a 3d morphable model to edges: A comparison between hard and soft correspondences. In *Asian Conference on Computer Vision Workshops*, pages 377–391, Cham, 2017. Springer International Publishing. 3

[7] Thabo Beeler, Bernd Bickel, Paul Beardsley, Bob Sumner, and Markus Gross. High-quality single-shot capture of facial geometry. *ACM Trans. Graph.*, 29(4), 2010. 2

[8] Thabo Beeler, Fabian Hahn, Derek Bradley, Bernd Bickel, Paul Beardsley, Craig Gotsman, Robert W. Sumner, and Markus Gross. High-quality passive facial performance capture using anchor frames. In *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*, pages 75:1–75:10, New York, NY, USA, 2011. ACM. 1, 2, 3

[9] Volker Blanz, Curzio Basso, Tomaso Poggio, and Thomas Vetter. Reanimating faces in images and video. In *Compututer Graphics Forum*, volume 22, pages 641–650. Wiley Online Library, 2003. 2

[10] Volker Blanz, Sami Romdhani, and Thomas Vetter. Face identification across different poses and illuminations with a 3d morphable model. In *Proc. International Conference on Automatic Face and Gesture Recognition*, pages 202–207. IEEE, 2002. 3

[11] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *ACM Transactions on Graphics (TOG)*, SIGGRAPH '99, 1999. 2, 3

[12] 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. 13

[13] Timo Bolkart and Stefanie Wuhrer. A groupwise multilinear correspondence optimization for 3D faces. In *Proc. International Conference on Computer Vision (ICCV)*, pages 3604–3612, 2015. 3

[14] George Borshukov, Dan Piponi, Oystein Larsen, John Peter Lewis, and Christina Tempelaar-Lietz. Universal capture–image-based facial animation. *The Matrix Reloaded. SIGGRAPH*, 2003. 3

[15] George Borshukov, Dan Piponi, Oystein Larsen, John P Lewis, and Christina Tempelaar-Lietz. Universal capture-image-based facial animation for" the matrix reloaded". In *ACM Siggraph 2005 Courses*, pages 16–es. 2005. 3

[16] Feng-Ju Chang, Anh Tuan Tran, Tal Hassner, Iacopo Masi, Ram Nevatia, and Gerard Medioni. Expnet: Landmark-free, deep, 3D facial expressions. In *Proc. International Conference on Automatic Face and Gesture Recognition*, pages 122–129, 2018. 3

[17] Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. Acquiring the reflectance field of a human face. In *Proc. Conference on Computer graphics and Interactive Techniques*, pages 145–156. ACM Press/Addison-Wesley Publishing Co., 2000. 5

[18] Bernhard Egger, William A. P. Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhöfer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, Christian Theobalt, Volker Blanz, and Thomas Vetter. 3D morphable face models - past, present and future. *ACM Trans. Graph.*, 2020. 2, 3

[19] Carlos Hernández Esteban and Francis Schmitt. Silhouette and stereo fusion for 3d object modeling. *Computer Vision and Image Understanding*, 96(3):367–392, 2004. 2

[20] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3D face model from in-the-wild images. *ACM Transactions on Graphics (ToG), Proc. SIGGRAPH*, 40(4):88:1–88:13, 2021. 3

[21] Yasutaka Furukawa. *High-fidelity image-based modeling*. University of Illinois at Urbana-Champaign, 2008. 2

[22] G. Fyffe, K. Nagano, L. Huynh, S. Saito, J. Busch, A. Jones, H. Li, and P. Debevec. Multi-view stereo on consistent face topology. *Compututer Graphics Forum*, 36(2):295–309, May 2017. 2, 3

[23] Kyle Genova, Forrester Cole, Aaron Maschinot, Aaron Sarna, Daniel Vlasic, and William T. Freeman. Unsupervised training for 3D morphable model regression. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8377–8386, 2018. 3

[24] 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. 1, 2, 5, 12

[25] Michael Goesele, Brian Curless, and Steven M Seitz. Multi-view stereo revisited. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 2402–2409. IEEE, 2006. 2

[26] Paulo Gotardo, Jérémy Riviere, Derek Bradley, Abhijeet Ghosh, and Thabo Beeler. Practical dynamic facial appearance modeling and acquisition. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)*, 37(6):232:1–232:13, 2018. 1

[27] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[29] Liwen Hu, Shunsuke Saito, Lingyu Wei, Koki Nagano, Jaewoo Seo, Jens Fursund, Iman Sadeghi, Carrie Sun, Yen-

Chun Chen, and Hao Li. Avatar digitization from a single image for real-time rendering. *ACM Trans. Graph.*, 36(6), 2017. 1

[30] Alexandru Eugen Ichim, Sofien Bouaziz, and Mark Pauly. Dynamic 3d avatar creation from hand-held video input. *ACM Transactions on Graphics (ToG)*, 34(4):1–14, 2015. 1

[31] Sunghoon Im, Hyowon Ha, Hae-Gon Jeon, Stephen Lin, and In So Kweon. Deep depth from uncalibrated small motion clip. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 4

[32] Sunghoon Im, Hae-Gon Jeon, Stephen Lin, and In-So Kweon. Dpsnet: End-to-end deep plane sweep stereo. In *7th International Conference on Learning Representations, ICLR 2019*. International Conference on Learning Representations, ICLR, 2019. 2

[33] Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. Learnable triangulation of human pose. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7718–7727, 2019. 4, 5

[34] Abhishek Kar, Christian Häne, and Jitendra Malik. Learning a multi-view stereo machine. In *Advances in neural information processing systems*, pages 365–376, 2017. 2

[35] Vladimir Kolmogorov and Ramin Zabih. Multi-camera scene reconstruction via graph cuts. In *European conference on computer vision*, pages 82–96. Springer, 2002. 2

[36] Marc Levoy, Kari Pulli, Brian Curless, Szymon Rusinkiewicz, David Koller, Lucas Pereira, Matt Ginzton, Sean Anderson, James Davis, Jeremy Ginsberg, et al. The digital michelangelo project: 3D scanning of large statues. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 131–144, 2000. 2

[37] Hao Li, Bart Adams, Leonidas J Guibas, and Mark Pauly. Robust single-view geometry and motion reconstruction. *ACM Transactions on Graphics (ToG)*, 28(5):1–10, 2009. 2, 3

[38] Jiaman Li, Zhengfei Kuang, Yajie Zhao, Mingming He, Karl Bladin, and Hao Li. Dynamic facial asset and rig generation from a single scan. *ACM Transactions on Graphics (TOG)*, 39(6):1–18, 2020. 2

[39] Ruilong Li, Karl Bladin, Yajie Zhao, Chinmay Chinara, Owen Ingraham, Pengda Xiang, Xinglei Ren, Pratusha Prasad, Bipin Kishore, Jun Xing, and Hao Li. Learning formation of physically-based face attributes. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 5

[40] 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. 3, 13

[41] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *arXiv preprint arXiv:1906.07751*, 2019. 1

[42] 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. 1, 2, 5, 12

[43] Koki Nagano, Jaewoo Seo, Jun Xing, Lingyu Wei, Zimo Li, Shunsuke Saito, Aviral Agarwal, Jens Fursund, Hao Li, Richard Roberts, et al. pagan: real-time avatars using dynamic textures. In *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)*, page 258. ACM, 2018. 1

[44] J-P Pons, Renaud Keriven, and Olivier Faugeras. Modelling dynamic scenes by registering multi-view image sequences. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 822–827. IEEE, 2005. 2

[45] 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, 5, 8, 12

[46] Elad Richardson, Matan Sela, and Ron Kimmel. 3D face reconstruction by learning from synthetic data. In *Proc. IEEE International Conference on 3D Vision (3DV)*, pages 460–469, 2016. 3

[47] Jérémy Riviere, Paulo Gotardo, Derek Bradley, Abhijeet Ghosh, and Thabo Beeler. Single-shot high-quality facial geometry and skin appearance capture. 2020. 1

[48] Olaf Ronneberger, Philipp Fischer, and Thomas 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. 5

[49] Augusto Salazar, Stefanie Wuhrer, Chang Shu, and Flavio Prieto. Fully automatic expression-invariant face correspondence. *Machine Vision and Applications*, 25(4):859–879, 2014. 2

[50] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael Black. Learning to regress 3D face shape and expression from an image without 3D supervision. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 7763–7772, 2019. 3

[51] Yeongho Seol, Wan-Chun Ma, and JP Lewis. Creating an actor-specific facial rig from performance capture. In *Proceedings of the 2016 Symposium on Digital Production*, pages 13–17, 2016. 1

[52] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhofer. Deepvoxels: Learning persistent 3d feature embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2437–2446, 2019. 2

[53] Olga Sorkine, Daniel Cohen-Or, Yaron Lipman, Marc Alexa, Christian Rössl, and H-P Seidel. Laplacian surface editing. In *Proceedings of the 2004 Eurographics/ACM SIGGRAPH symposium on Geometry processing*, pages 175–184, 2004. 5

[54] Ayush Tewari, Florian Bernard, Pablo Garrido, Gaurav Bharaj, Mohamed Elgharib, Hans-Peter Seidel, Patrick Pérez, Michael Zollhoefer, and Christian Theobalt. Fml: Face model learning from videos. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[55] Ayush Tewari, Michael Zollhoefer, Florian Bernard, Pablo Garrido, Hyeongwoo Kim, Patrick Perez, and Christian Theobalt. High-fidelity monocular face reconstruction based on an unsupervised model-based face autoencoder. *IEEE Transactions on Pattern Analysis and Machine Intelligence*,

pages 1–1, 2018. 6

[56] Ayush Tewari, Michael Zollhoefer, Pablo Garrido, Florian Bernard, Hyeongwoo Kim, Patrick Pérez, and Christian Theobalt. Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2549–2559, 2018. 6

[57] Ayush Tewari, Michael Zollhoefer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Theobalt Christian. MoFA: Model-based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction. In *Proc. International Conference on Computer Vision (ICCV)*, 2017. 1, 3

[58] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *CVPR*, 2016. 3

[59] Anh Tuan Tran, Tal Hassner, Iacopo Masi, Eran Paz, Yuval Nirkin, and Gérard G Medioni. Extreme 3d face reconstruction: Seeing through occlusions. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3935–3944, 2018. 3

[60] Luan Tran, Feng Liu, and Xiaoming Liu. Towards high-fidelity nonlinear 3d face morphable model. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[61] George Vogiatzis, Philip HS Torr, and Roberto Cipolla. Multi-view stereo via volumetric graph-cuts. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 391–398. IEEE, 2005. 2

[62] 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. 5, 6, 14

[63] 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. 5, 6, 15

[64] Fanzi Wu, Linchao Bao, Yajing Chen, Yonggen Ling, Yibing Song, Songnan Li, King Ngi Ngan, and Wei Liu. Mvf-net: Multi-view 3d face morphable model regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1, 2, 6

[65] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 767–783, 2018. 2

[66] Chao Zhang, William Smith, Arnaud Dessein, Nick Pears, and Hang Dai. Functional faces: Groupwise dense correspondence using functional maps. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

A. Appendix

Supplemental Video. Please see the video on the project page: https://tianyeli.github.io/tofu.

Additional Quantitative Results. Tab. 2 provides additional quantitative comparisons to other learning based methods, namely 3DMM regression and DFNRMVS [4]. Fig. 13 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 [4] (PP)	1.885	4.565
Our Method	**0.585**	**1.973**

Table 2: 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 [4] is higher than for the 3DMM regression, DFNRMVS [4] 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. 16, 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 [4] 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. 12 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. 20 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*.

Impact of Local Refinements. Fig. 14 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. 17 demonstrates the effect of each local refinement step. As shown in Fig. 17, the quality of the reconstructed mesh improves after each local stage, while the scan-to-mesh distance to the scan reduces. Note that details such as nose corners and lips gradually improve through the local stages.

More Ablation on Number of Views. Fig. 15 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. 21 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.

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. 19. 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. 19. 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 [24, 42], the results, shown in the paper, indicate that our system can be adapted to non-lightstage setups, e.g. capture system of CoMA [45]. 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

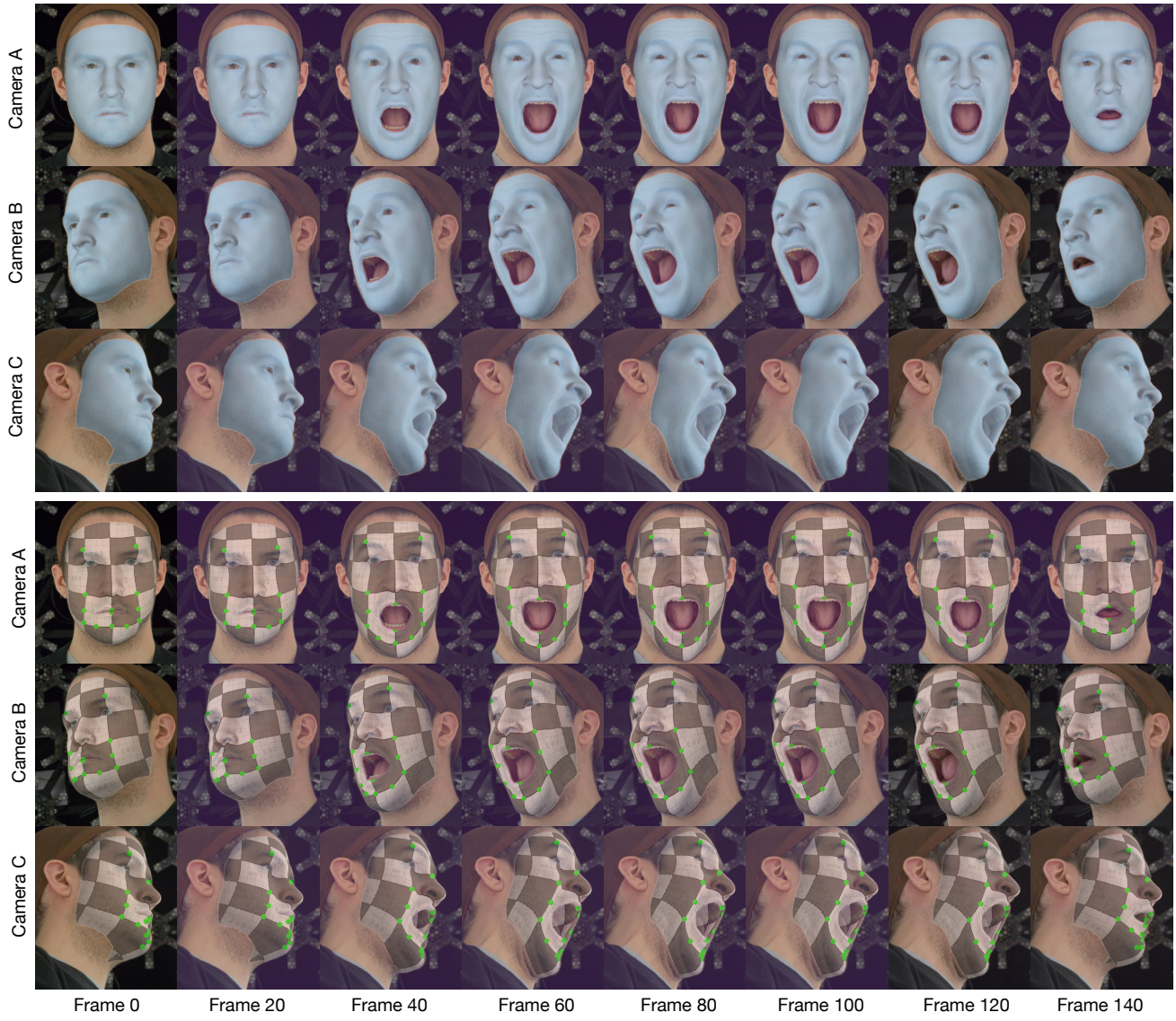

Figure 12: Base mesh reconstruction for a multi-view video sequence overlaid on the video frames. Our method captures the facial performance well. The result meshes are temporally stable and accurately align with the input images. Visualizing with a shared checkerboard texture indicates good tracking quality. Please see the *supplemental video* for better visualization.

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 [12, 40]. 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

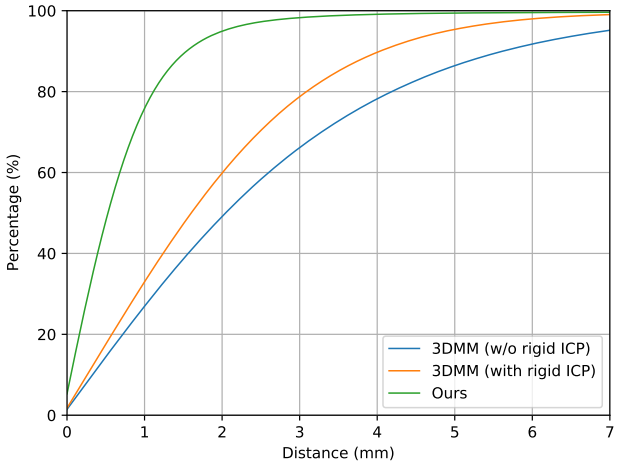

Figure 13: Quantitative evaluation by cumulative error curves for scan-to-mesh distances among learning based methods.

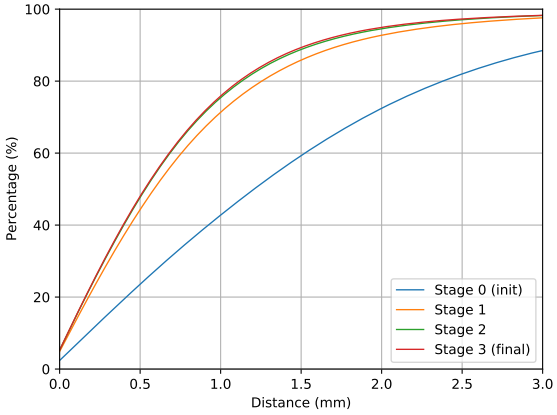

Figure 14: Quantitative evaluation by cumulative error curves for scan-to-mesh distances among local refinement stages.

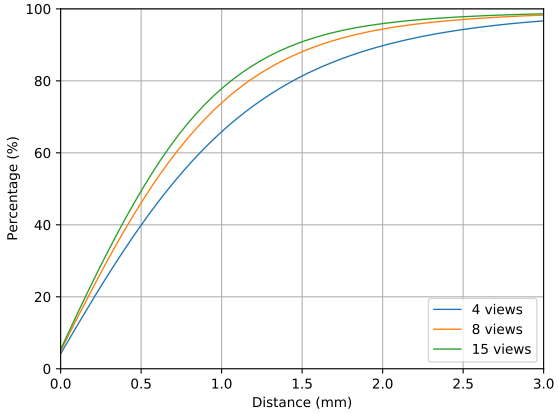

Figure 15: Quantitative evaluation by cumulative error curves for scan-to-mesh distances among various numbers of views.

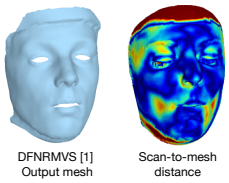

Figure 16: Example results from DFNRMVS [4].

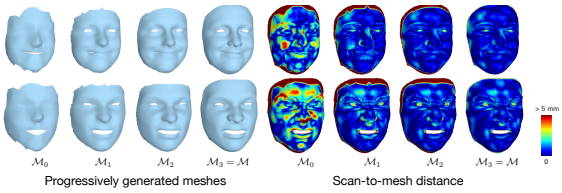

Figure 17: Inferred meshes for global stage $\mathcal{M}_0$ and after upsampling and refinement for each local stage $\mathcal{M}_i$ ($1 \leq i \leq 3$).

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. 12 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. 18. 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. [62]. 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

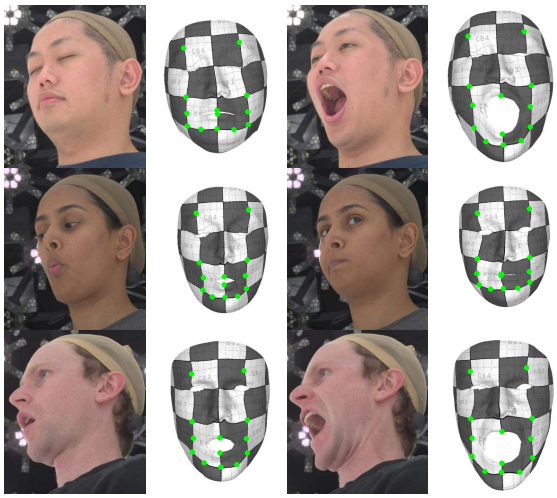

Figure 18: Visualization of cross-subject dense correspondence of the base meshes inferred by ToFu in a shared checkerboard texture.

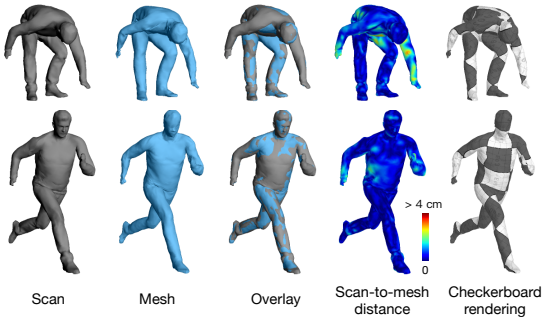

Figure 19: Our system can also infer clothed human body surfaces in consistent topology.

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 [63] 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.

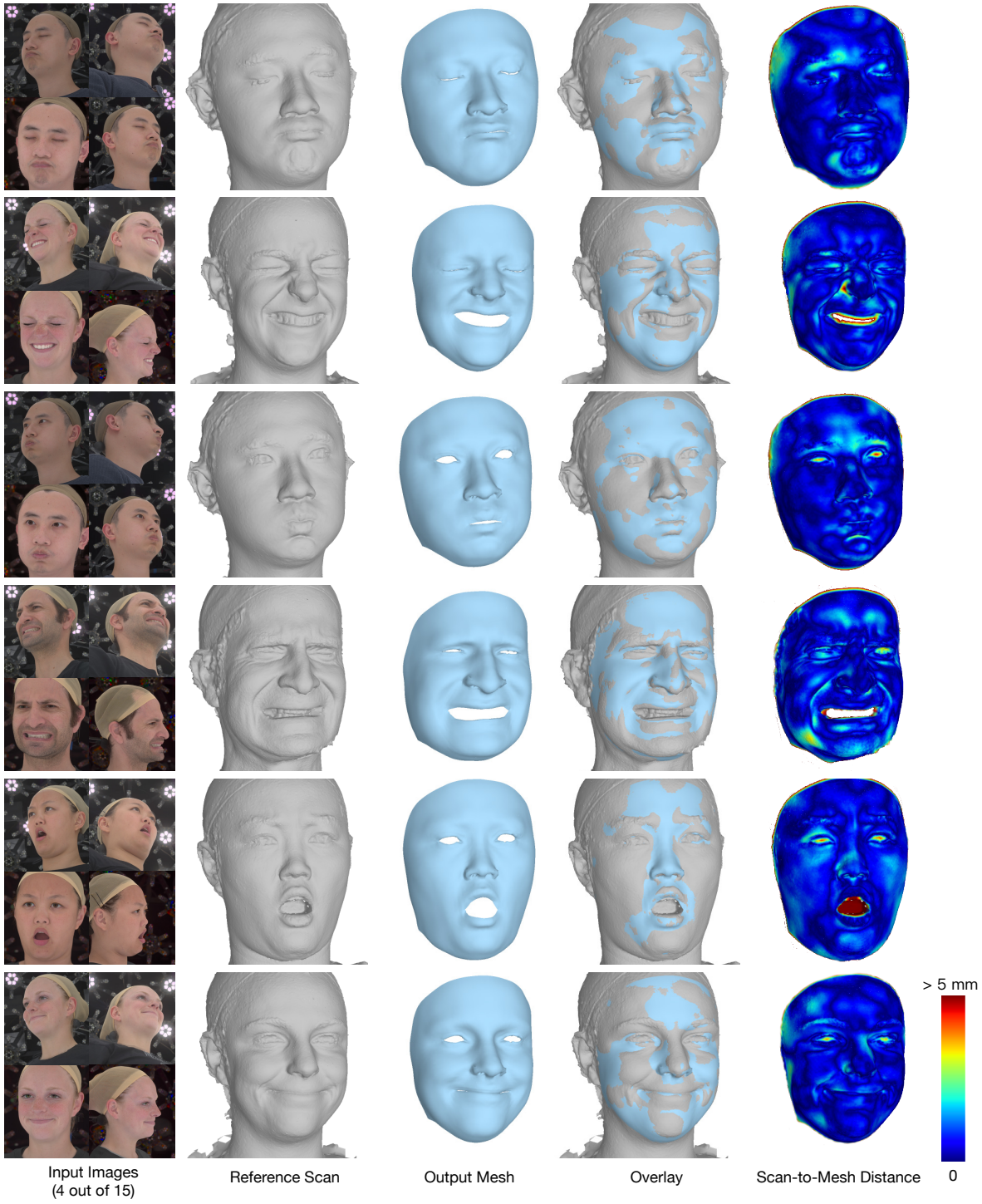

Figure 20: More results of reconstructed meshes in dense correspondence. The scan-to-mesh distance is visualized color coded on the reference scan, where red denotes an error above 5 millimeters.

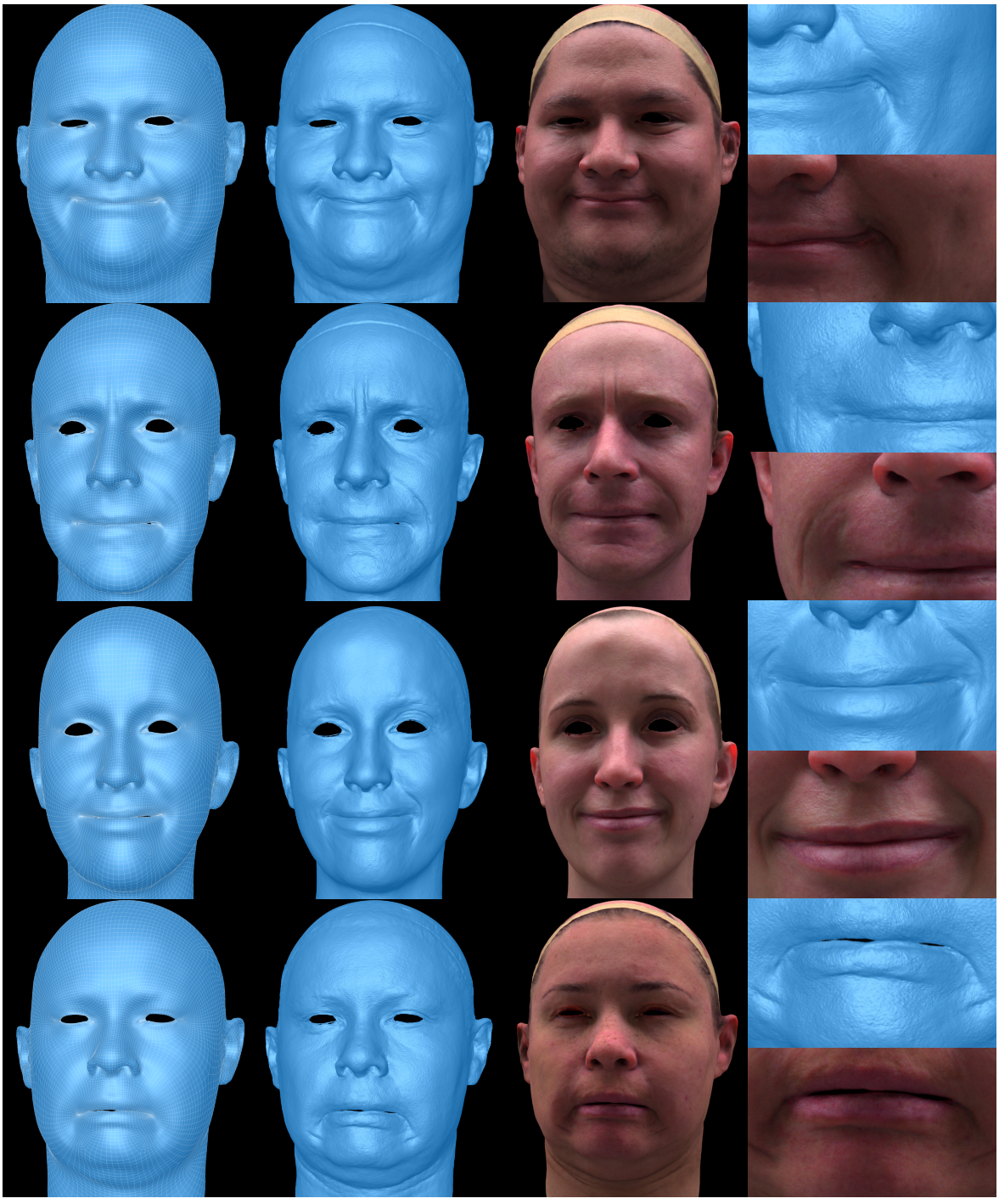

Figure 21: Our method can generate reliable base alignment meshes, on top of which a comprehensive face modeling pipeline can be built. Here we show more rendering with inferred normal displacements and additional albedo and specular maps.