Instant Multi-View Head Capture through Learnable Registration
Supplemental Material

Timo Bolkart[1] Tianye Li[2] Michael J. Black[1]
[1]MPI for Intelligent Systems, Tübingen [2]University of Southern California

Multi-view setup: TEMPEH infers 3D head meshes in correspondence from calibrated multi-view images. Specifically, we use the eight pairs of gray-scale stereo images of an active stereo camera system as input (see Sec. 4 of the paper for details). Figure 1 shows the 16 images for the first sample of the paper's teaser figure.

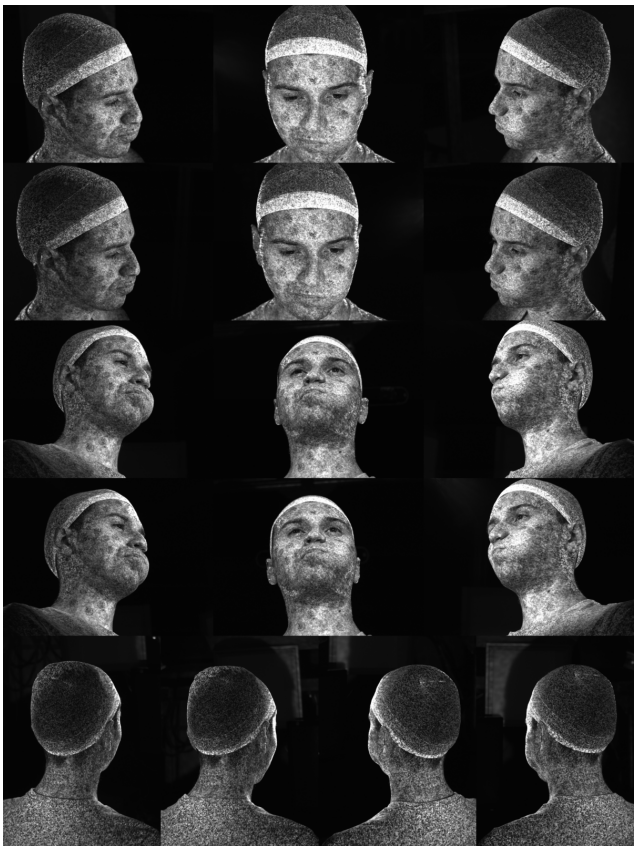

Figure 1. **Multi-view setup**. The 16 gray-scale stereo images (contrast enhanced for visualization) used as input to TEMPEH.

Head localization: The coarse head prediction stage localizes the head in the feature volume with a learnable spatial transformer. Figure 2 visualizes the spatial dimensions of the localized volume for different predicted heads of the coarse stage. This shows that the spatial transformer successfully localizes the head for different subjects in varying head poses.

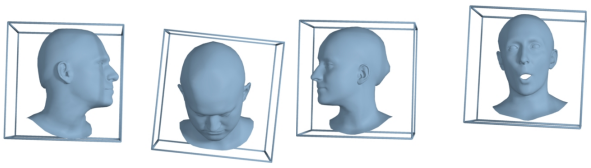

Figure 2. **Head localization**. Spatial dimensions of the localized feature volumes (blue box) for different predicted heads, visualized in the same global coordinate system.

Error masks: To analyze the accuracy in different head regions, reconstruction errors are reported individually for the face, scalp, and neck regions. For this, different regions are defined on a FLAME template mesh (see Figure 3), and each scan is then segmented based on the distance to the closest points in the surface of the reference registrations.

Test evaluation: Figure 4 provide the cumulative reconstruction errors for the FaMoS test data. TEMPEH predicts heads for the test images (subjects disjoint from the training subjects) with a lower error than previous state-of-the art, ToFu [4], and its variant without mesh hierarchy, ToFu+.

3DMM regressor comparisons: For the multi-view 3DMM regressor, each image is processed by a shared ResNet152 [2] to infer a 2048-dimensional feature vector for each view. The feature vectors are then fused across all 16 views by concatenating them in a fixed order. We experimented with other feature fusion variants such as using the mean across views, or the concatenated mean and variance, but these variants produced inferior results. Following DECA [1], a fully-connected layer with ReLU activations outputs a 1024-dimensional feature vector, followed by final linear layer to output FLAME parameters. We train the 3DMM regressor for 1 Million iterations with a vertex-to-vertex to the reference registrations, with a learning rate of 1e-3. We found that the 3DMM regressor is unable to reliably reconstruct 3D heads in our setting (see Fig. 5).

Registration quality: The training of TEMPEH minimizes

Method	Complete head			Face			Scalp			Neck		
	Median ↓	Mean ↓	Std ↓	Median ↓	Mean ↓	Std ↓	Median ↓	Mean ↓	Std ↓	Median ↓	Mean ↓	Std ↓
Ours (coarse)	0.80	1.61	3.86	0.67	0.85	1.31	0.84	2.31	5.59	1.11	1.68	2.34
Ours	0.17	0.30	0.97	0.14	0.23	1.10	0.16	0.24	0.39	0.24	0.53	1.46

Table 1. **Registration quality**. Reconstruction errors on the FaMoS training data. All errors are in millimeter.

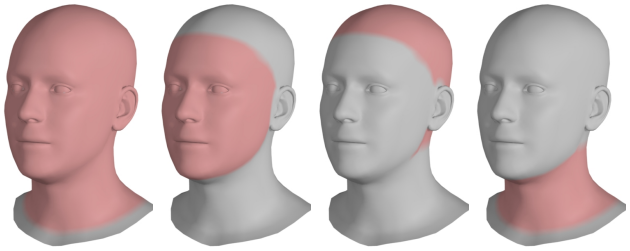

Figure 3. **Error masks**. Head regions (red) for quantitative evaluations. From left to right: *complete head*, *face*, *scalp*, and *neck*.

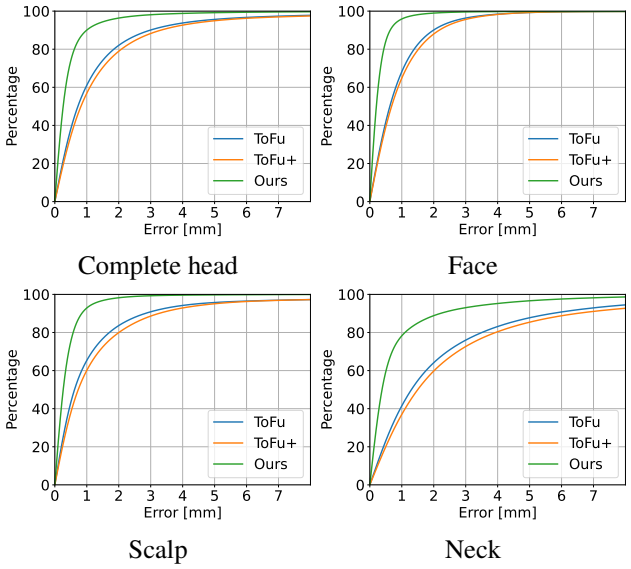

Figure 4. **Quantitative evaluation**. Cumulative plots of the reconstruction errors on the FaMoS test data.

the distance to multi-view stereo (MVS) scans, hence it effectively registers the scans. For completeness, we also report the registration errors in Table 1. For this, we predict all heads from the training images, and compute the point-to-surface distance for all MVS scan points.

Ablation experiments: Figure 6 shows additional ablation results. While the model variants without head localization (Ours w/o head localization) or with a hierarchical architecture (Ours hierarchical) produce reconstructions with low distance to the reference scans, they reconstruct the lip region with lower fidelity than the final model.

Failures: TEMPEH's coarse stage reconstruction can fail

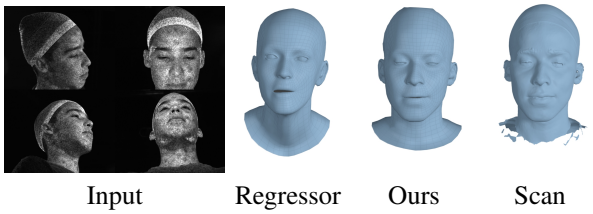

Figure 5. **3DMM regressor comparison**. For multi-view input (left: 4 of 16 views), the 3DMM regressor regressor (second column) is unable to faithfully reconstruct the identity face shape, while TEMPEH (third column) closely resembles the reference scan (right).

under large occlusions due to extreme head poses (see Figure 7). We found empirically that training the coarse stage for 250K more iterations improves the quality of the reconstructed head meshes for such extreme head poses.

Computational requirements: TEMPEH and the baseline models are trained/evaluated on a computing unit with a single NVIDIA A100-SXM 80 GB GPU and 16 CPU cores. Training TEMPEH/ToFu/ToFu+ allocates 26/4/14 GB GPU memory for the coarse stage, 37/34/37 GB for refinement, and up to 21 GB RAM. GPU memory is mainly allocated in the volumetric feature sampling and the probability volume prediction for the 2 (batch) $\times 5023$ local grids of size 8^3. Training TEMPEH takes 6 days ($3.5/2.5$ days for coarse / refinement). Inference for TEMPEH/ToFu/ToFu+ allocates 6/4/6 GB GPU memory for the coarse stage and 10/6/8 GB for refinement.

Running time evaluation: TEMPEH targets the typical two-step process of reconstructing 3D meshes in correspondence, MVS, followed by non-rigid registration. This pipeline takes $\geq$ 10 minutes per mesh (Tab. 1 [4]), while TEMPEH takes 0.27s. While ToFu/ToFu+ is even faster with 0.16/0.18s, TEMPEH reconstructs 3D heads with a 64% lower error. The time difference between ToFu/ToFu+ and TEMPEH is mainly due to the visibility computation in the surface-aware feature fusion. TEMPEH w/ naïve feature fusion requires 0.17s, comparable to ToFu/ToFu+. The coarse model inference accounts for about 0.03s for all models. The fast inference speed is due to downsampling of the input images, and due to parallelization. Specifically, the feature extraction is parallelized across images (stacked across the batch dimension), while feature sampling & aggregation, head inference, and mesh refinement are paral-

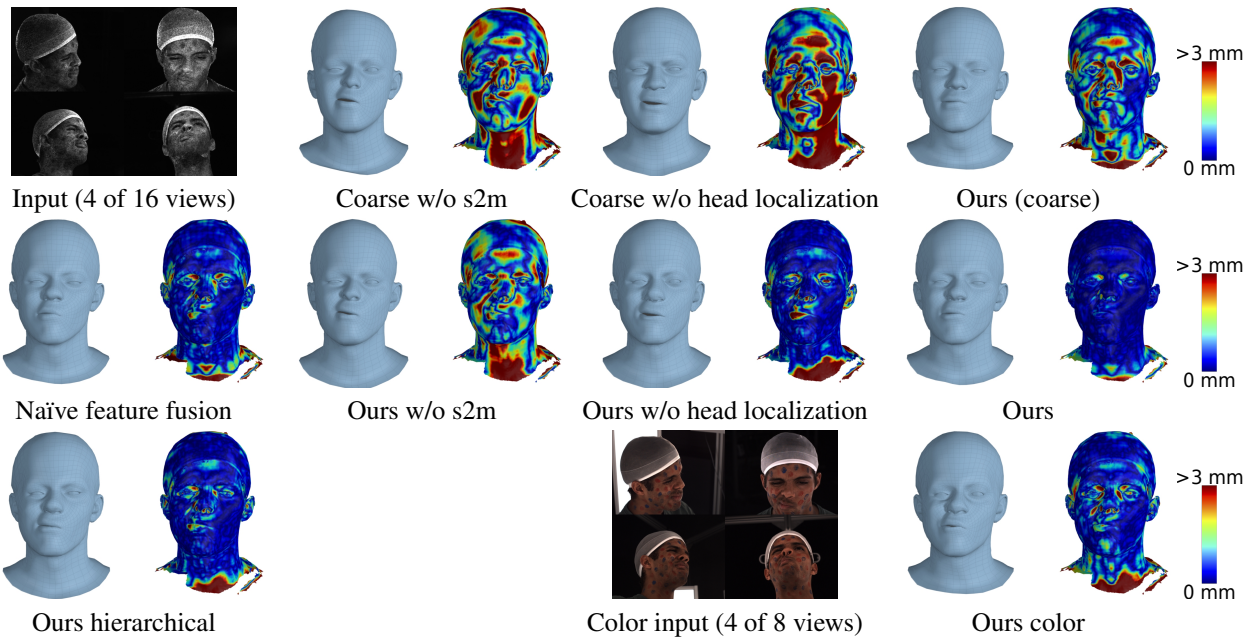

Figure 6. **Additional ablation experiments**. For each model variant, we show the reconstructed mesh (left) and the color coded point-to-surface distance (right) between reference scan and reconstructed mesh as heatmap on the scan's surface (red means ≥ 3 millimeter).

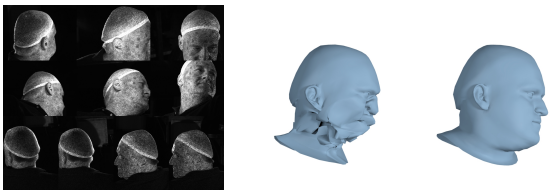

Figure 7. **Failures**. For input images (left: 10 of 16 views) where the face is occluded in most views, the coarse stage reconstruction can fail (middle), resulting in poorly estimated head meshes. Longer training of the coarse stage can improve the reconstruction performance for such extreme cases (right).

lelized across all points.

Data diversity: FaMoS data are female/male: 52/41; age 18-34: 65, 35-50: 14, 51-69: 13, 70+: 1; Middle-Eastern: 6, South American: 10, Asian: 24, Pacific Ocean: 1, African: 3, European: 49. We provide self-identified ethnicity labels as provided by each participant with the dataset.

Model architecture: TEMPEH uses volumetric features to localize, infer and then refine the output mesh. These features are extracted from the input images with two separate 2D feature extraction networks F_{img}, one for coarse head prediction (Section 3.1) and one for head refinement (Section 3.2). Both networks use a fully-convolutional U-Net [5] architecture with a ResNet34 [2] backbone. Both feature networks take downsampled images as input (i.e., images of size w = 200, h = 150 for the coarse stage, and w = 400, h = 300 for the refinement stage), and output a feature map $\mathcal{F}$ with the same spatial resolution as the image, with a feature dimension of 8. We empirically found that adding two additional skip connections for the feature networks compared to ToFu's implementation improved the reconstruction performance of the refinement network. For a fair comparison to ToFu and ToFu+, we use the same feature extractor networks with added skip connections for all models.

The reconstruction networks in coarse and refinement stages, F_{rec} and F_{ref}, respectively, are both 3D U-Nets [3]. Similar to ToFu, the coarse stage reconstruction network F_{rec} has five down- and upsampling blocks, with a slight modification of the third last and second last convolution blocks, which output 64 and 128 channels (instead of 32 for ToFu). The refinement stage reconstruction network F_{ref} follows a similar structure, but with three down- and upsampling layers, same as ToFu.

References

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

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 1, 3

[3] Karim Iskakov, Egor Burkov, Victor S. Lempitsky, and Yury Malkov. Learnable triangulation of human pose. In *International Conference on Computer Vision (ICCV)*, pages 7717–7726, 2019. 3

[4] Tianye Li, Shichen Liu, Timo Bolkart, Jiayi Liu, Hao Li, and Yajie Zhao. Topologically consistent multi-view face inference using volumetric sampling. In *International Conference on Computer Vision (ICCV)*, pages 3824–3834, 2021. 1, 2

[5] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241, 2015. 3